明 細 書

情報抽出システム

技術分野

[0001] 本発明は、テキストから事物に関して書かれた事実や意見等の記述内容を抽出する情報抽出システムに関する。

背景技術

[0002] 従来の情報抽出システムとしては、テキストからキーワードを抽出するもの、固有名や数値表現等を抽出するもの、5W1H等の事実に関する情報を抽出するもの、意見や評判を抽出するもの等が知られている。狭義の情報抽出は、非特許文献1に紹介されているように、テキストの中心的な情報を抽出するもので、特定の分野のテキストを対象に抽出すべき情報のテンプレート(またはフレーム)を用意しておき、該当する情報を抽出するのがその典型である。一方、近年はテキスト中の意見や評判を抽出しようとする研究が行われている。例えば、特許文献1は利用者が指定した物に関する意見を文書集合中から抽出するものである。

特許文献1:特開2003－203136号公報

非特許文献1:長尾他著『自然言語処理』岩波書店(pp. 438－441, 1996)

発明の開示

発明が解決しようとする課題

[0003] しかしながら、特許文献1のような従来の意見情報抽出システムでは事物に関する意見を抽出することは可能であるが、事物に関して書かれた事実や意見の観点と記述を対応付けて抽出する事が出来ないという課題がある。

[0004] 本発明は、かかる点に鑑みてなされたものであり、その第1の目的は、テキスト中に表現された事物に関する事実や意見などの記述内容を、事実や意見の観点と記述を対応付けて抽出する情報抽出システムを提供することである。

[0005] 本発明の第2の目的は、前記事実や意見などの記述内容を抽出するに当たって、事実や意見の対応付けや関連性の比較が容易に行える形に整理して抽出することができる情報抽出システムを提供することである。

課題を解決するための手段

[0006] 上記課題を解決するため、本発明の情報抽出システムは、テキストを入力する入力部と、テキストに記述された表現の観点とその観点に関する記述の組を特定するための観点・記述抽出規則を格納する観点・記述抽出規則格納部と、前記テキスト中の文字列の統語的属性または意味的属性の少なくとも一方の属性から、前記観点・記述抽出規則を用いて観点とその記述の組を対応付け、それらを識別するための識別情報を付与した要素メタデータとして抽出する観点・記述抽出部と、前記観点・記述抽出部が抽出した要素メタデータを格納するメタデータ格納部を有する構成をとる。

[0007] この構成によれば、テキスト中に表現された事物に関する事実や意見などの記述内容を、観点と記述の組として構成し、事実や意見を対応付けて抽出することができる。さらに、その後の処理で抽出された事実や意見について、関連性の比較が容易に行える形に整理することができる。

発明の効果

[0008] 以上説明したように、本発明の情報抽出システムは、テキストに記述された表現の観点とその観点に関する記述の組を特定するための観点・記述抽出規則を用いて観点とその記述の組を対応付けて抽出することにより、テキスト中に表現された事物に関する事実や意見の記述内容を、観点と記述の組として対応付けて抽出することができるという効果を有する。

[0009] 本発明の上記目的及び利点は添付図面を参照して説明される、以下の実施例によってより一層明らかになるであろう。

図面の簡単な説明

[0010] [図1]本発明の実施の形態1に係る情報抽出システムの構成を示すブロック図

[図2]実施の形態1に係る情報抽出システムにおける、テキストから要素メタデータを抽出するまでの一連の処理の流れを示す説明図

[図3]実施の形態1に係る情報抽出システムにおける、観点・記述抽出規則と規則の構成要素定義の例を示す図

[図4]実施の形態1に係る情報抽出システムにおける、統合メタデータの例を示す図

[図5]本発明の実施の形態2に係る情報抽出システムの構成を示すブロック図

[図6]実施の形態2に係る情報抽出システムにおける、入力されたテキストと意味属性を付与されたテキストの例を示す図

[図7]実施の形態2に係る情報抽出システムにおける、意味属性付与規則の例と意味属性付与規則構成要素定義の例を示す図

[図8]実施の形態2に係る情報抽出システムにおける、意味属性付きテキストの例と観点・記述認定例を示す図

[図9]実施の形態2に係る情報抽出システムにおける、観点・記述抽出規則と規則の構成要素定義の例を示す図

[図10]実施の形態2に係る情報抽出システムにおける、要素メタデータ抽出結果の例を示す図

[図11]実施の形態2に係る情報抽出システムにおける、統合メタデータの例を示す図

[図12]本発明の実施の形態3に係る情報抽出システムの構成を示すブロック図

[図13]実施の形態3に係る情報抽出システムにおける、観点・記述の認定結果と要素メタデータの抽出結果を示す図

[図14]実施の形態3に係る情報抽出システムにおける、話題事物推定規則と話題事物推定規則構成要素定義の例を示す図

[図15]実施の形態3に係る情報抽出システムにおける、推定した話題事物の例を示す図

[図16]実施の形態3に係る情報抽出システムにおける、統合メタデータの例を示す図

[図17]実施の形態3に係る情報抽出システムにおける、メタデータ出力形式の例を示す図

[図18]本発明の実施の形態4に係る情報抽出システムの構成を示すブロック図

[図19]実施の形態4に係る情報抽出システムにおける、テキストのソース情報、ユーザ情報の例と意味属性付きソース情報、意味属性付きユーザ情報の例を示す図

[図20]実施の形態4に係る情報抽出システムにおける、ソース情報意味属性付与規則、ユーザ意味属性付与規則の例を示す図

[図21]実施の形態4に係る情報抽出システムにおける、ソース観点・記述抽出規則、ユーザ観点・記述抽出規則の例を示す図

[図22]実施の形態4に係る情報抽出システムにおける、ソースメタデータ抽出結果、ユーザメタデータ抽出結果の例を示す図

[図23]実施の形態4に係る情報抽出システムにおける、客観性・信頼性判定規則と客観性・信頼性判定規則構成要素定義の例を示す図

[図24]実施の形態4に係る情報抽出システムにおける、テキストの例と意味属性付きテキストの例を示す図

[図25]実施の形態4に係る情報抽出システムにおける、観点・記述抽出規則例と観点・記述構成要素定義例を示す図

[図26]実施の形態4に係る情報抽出システムにおける、要素メタデータ抽出結果の例を示す図

[図27]実施の形態4に係る情報抽出システムにおける、客観性・信頼性判定結果の例を示す図

[図28]実施の形態4に係る情報抽出システムにおける、メタデータ統合結果の例を示す図

[図29]実施の形態4に係る情報抽出システムにおける、メタデータ出力形式の例を示す図

符号の説明

[0011] 100, 200, 300, 400 情報抽出システム

102 入力部

106 メタデータ照合部

108 メタデータ統合部

110 メタデータ格納部

120 観点・記述抽出部

122 観点・記述抽出規則格納部

202 属性付与部

204 意味属性付与規則格納部

206 意味属性付きテキスト格納部

302 ユーザ要求処理部

304 メタデータ出力形式生成部

306 メタデータ出力部

310 話題事物推定部

312 話題事物推定規則格納部

412 客観性・信頼性判定部

414 客観性・信頼性判定規則格納部

発明を実施するための最良の形態

[0012] 以下、本発明の実施の形態を、図面を参照して詳細に説明する。

[0013] (実施の形態1)

図1は、本発明の実施の形態1に係る情報抽出システムの構成を示すブロック図である。本実施の形態の情報抽出システム100は、入力されたテキスト中に表現された事物に関する事実や意見などの記述内容を、観点や記述の組として構成し、事実や意見の対応付けや関連性の比較が容易に行える形に整理して抽出するためのシステムである。情報抽出システム100は、テキストが入力される入力部102と、テキストに記述された表現の観点とその観点に関する記述の組を特定するための観点・記述抽出規則を格納する観点・記述抽出規則格納部122と、前記テキスト中の文字列の統語的属性から、前記観点・記述抽出規則を用いて観点とその記述の組を対応付け、それらを識別するための識別情報を付与した要素メタデータとして抽出する観点・記述抽出部120と、観点・記述抽出部120が抽出した要素メタデータの観点間、記述間をそれぞれ照合し、要素メタデータの関連性を推定するメタデータ照合部106と、前記推定された関連性に基づいて関連性のある要素メタデータを統合するメタデータ統合部108と、メタデータ統合部108により統合された要素メタデータである統合メタデータを格納するメタデータ格納部110とを有する。

[0014] なお、情報抽出システム100のハードウエア構成は、任意であって、特に限定されない。たとえば、情報抽出システム100は、CPUや記憶装置(ROM、RAM、ハードディスクその他各種記憶媒体)を備えたコンピュータによって実現される。このように情報抽出システム100がコンピュータによって実現される場合は、この情報抽出システム100の動作を記述したプログラムをCPUが実行することによって所定の動作を

行う。

[0015] この情報抽出システム100では、まず、入力部102で入力されたテキストを受け取る。観点・記述抽出規則格納部122には、テキストに書かれた表現の観点とその観点に関する記述の組を特定するための観点・記述抽出規則が格納されている。観点・記述抽出部120は、観点・記述抽出規則格納部122に格納された、観点・記述抽出規則を参照し、前記テキスト中の文字列の統語的属性から、事物に関して記述された内容を観点とその記述の組として対応付ける。次に、対応づけられた観点とその記述の組にそれらを識別するための識別情報である要素メタデータIDを付与した要素メタデータとして抽出する。そして、メタデータ照合部106が、抽出された要素メタデータの観点間、記述間をそれぞれ比較・照合し、関連性を推定する。さらにメタデータ統合部108が、メタデータ照合部106の推定した関連性に基づいて関連性のある要素メタデータを統合し、統合メタデータとしてメタデータ格納部110に格納する。

[0016] ここで、メタデータとは、一般にコンテンツの内容や書誌事項等のコンテンツに関する情報を示すデータのことである。本発明では、テキスト中に表現された事物に関する事実や意見などのコンテンツの内容に関する記述内容を、観点と記述の組として構成したものをメタデータの基本単位とみなし、特に要素メタデータとよぶ。上記事実や意見という言葉の「事実」とは、誰が見ても客観的に同じであることが認定される事柄を意味し、例えばものの名前(固有名称を含む)や日時、或いは数量といったものを指す。「意見」とは、それぞれの事物に対して個々人がどのように考えたり感じたり評価したりするかといった見解を意味し、例えば重い、軽い、熱い、不十分だといったものを指す。「観点」とは、事物に関する事実や意見が、事物のどのような点に着目して、あるいはどのような見地から述べられているかということを意味する。また、「記述」とは、上記観点から具体的にどのような表現でテキスト中に言い表されているかということを意味する。ただし、要素メタデータを構成する観点と記述は、テキスト中に一方しか表現されない場合もある。なお、1つの観点に対して複数の記述がある場合は、1つの観点に対して複数の記述を抽出する。また、要素メタデータには、観点と記述の組だけでなく、それらの属性や話題等の関連情報も含めてよいこととする。また、複数の要素メタデータの観点や記述やそれらの関連情報のうち、関連するものを統合

した要素メタデータを統合メタデータとよぶ。

[0017] 要素メタデータは、識別情報である要素メタデータIDを付与されることとする。要素メタデータIDは、要素メタデータの出現したテキストと、個々の要素メタデータを識別するために個々の要素メタデータに付与される要素メタデータの識別情報である。また、文字列の統語的属性とは、文字列の構文的機能に関する属性であり、少なくとも品詞分類情報、または、文字列表記に関する情報のいずれかで指定されることとする。文字列表記に関する情報は、一部の語の境界の認定に用いられるもので、例えば字種を文字列表記に関する情報として用いることで、構文解析を実施していないテキストであっても、名詞連続と助詞の区切りの認定等の簡易な解析を行うことができる。

[0018] 次いで、上記構成を有する情報抽出システム100について、具体例を用いてより詳細に説明する。図2は、入力されたテキストから事物に関して表現された事実や意見などの内容を要素メタデータとして抽出するまでの一連の処理の概要を示す説明図である。図2では、図2(a)に入力テキスト例、図2(b)に観点・記述認定例、図2(c)に要素メタデータ抽出結果例をそれぞれ示す。

[0019] まず、観点・記述抽出部120は、観点・記述抽出規則格納部122に記憶された観点・記述抽出規則を参照し、入力部102から入力されたテキスト内の文字列が観点・記述抽出規則のパタンで指定された統語的属性を有するかどうか調べる。観点・記述抽出規則と規則の構成要素定義の例を図3に示す。ここで、規則の構成要素定義とは、規則中でパタンなどの記述に用いる文字列を構成要素としてあらかじめ定義しておくもので、規則中では構成要素名を記述すれば、その構成要素名で定義された文字列に相当するものとみなす。構成要素名の定義方法は、構成要素名と文字列あるいは文字列パタンのリストの対応付けが可能であれば特に限定されない。例えば、構成要素名と対応する文字列あるいは文字列パタンのリストを1ファイルに記述してもよいし、対応する文字列あるいは文字列パタンのリストは別の複数ファイルに記述しても構わない。なお、これ以降の規則例で同様の構成要素を用いる場合は定義を省略する。各規則には観点・記述を抽出するためのパタンと、パタン中で観点、記述に該当する箇所が示されている。

[0020] 図3(a)に示す観点・記述抽出規則は、文字列の統語的属性を用いて観点・記述を抽出するための規則である。観点・記述抽出規則のパタンには、観点・記述に相当する文字列またはその周辺の文字列の統語的属性が、文字列表記または品詞分類で指定されている。統語的属性を文字列表記で指定する場合は、規則のパタン中に、「は」のような文字列や、[がも](「が」「も」のいずれか、の意」のような文字列を含む正規表現のパタンとして記述するか、または、「漢字/平仮名連続1」のようにあらかじめ定義された構成要素名で指定する。統合的属性を品詞分類で指定する場合は、品詞分類名に対応する構成要素名を例えば「形容動詞語尾1」「形容詞語尾1」のようにあらかじめ定義しておき、定義された構成要素名を指定する。

[0021] なお、文字列の統語的属性の指定方法として、上記説明では文字列表記と品詞分類を用いたが、本発明はこれらに限定されるものではなく、他に例えば、構文的関係を用いても構わない。また、文字列表記や品詞分類を用いる場合も、それらの指定方法は上記の方法に限定されるものではなく、他の方法であっても構わない。また、統語的属性のかわりに意味的属性を用いて指定してもよいし、統語的属性と意味的属性の両方を指定してもよいし、さらにこれらに加えて統計的属性等の他の属性を指定しても構わない。また、規則を適用する条件を、上記説明では規則パタンのみで指定したが、パタンの一部に関する制約を別途指定してもよいし、パタン以外で指定しても構わない。

[0022] また、図3(a)において、パタン中で観点や記述に該当する箇所は"()"でマークされており、マークされた部分は先頭から順に＄1、＄2、…と参照される。例えば、規則1の場合は、＜「は」＞＜漢字／平仮名連続1＞＜「が」または「も」＞、＜英数字連続1＞、＜「と」＞、＜＜漢字／平仮名連続1＞、＜形容詞語尾1＞がこの順でテキスト中に出現した場合、規則のパタンと一致する。テキスト中のこのパタンに相当する文字列で、パタン中の、最初の"()"で括られた＜漢字／平仮名連続1＞に相当する部分が＄1として参照される。また、2番目の"()"で括られた＜英数字連続1＞に相当する部分が＄2として参照され、3番目の"()"で括られた＜漢字／平仮名連続1＞＜形容詞語尾1＞に相当する部分が＄3として参照される。規則にしたがって、＄1で参照される部分は観点、＄2、＄3で参照される部分は記述として抽出される。なお、

規則の記法は上記に限定されるものではなく、他の記法を用いても構わない。

[0023] 図3(a)の規則1を図2(a)のテキスト1に適用する場合、1文目の"開口部"が観点、"30cm"と"かなり大きい"が記述に該当する。図2(b)の観点・記述認定例はテキスト内の観点・記述対に、識別用の観点・記述対ID番号を与え、観点の表現の始まりと終わりを＜VIEW(観点・記述対の番号)＞…＜／VIEW(観点・記述対の番号)＞、記述の表現の始まりと終わりを＜DESC(観点・記述対の番号)＞…＜／DESC(観点・記述対の番号)＞でマークしたものである。なお、観点・記述対ID番号の与え方は観点・記述対を一意に特定できるものであれば特に限定されるものではない。例えば、テキストの識別情報とテキスト内での観点・記述対の番号を組み合わせたものでもよい。

[0024] なお、例えば「容量が20リットルと大きい」のように1つの観点(この例では「容量」)に対して、「20リットル」「大きい」と複数の記述がある場合には、これらを同じ観点に対する異なる2つの記述として認定する。本発明の観点と記述の抽出規則例では、同じ観点に対して複数の異なる記述を認定する場合、これらの記述を記号'||'を用いて例えば'$1||$2'(ただし、$1，$2は記述)のように示す。

[0025] 一方、例えば、「容量が旅行用には小さい」のように1つの観点(この例では「容量」)に対して、用途が「旅行用」に限定された場合は「小さい」というように、記述間に限定的な関係がある場合は、複数の記述(この例では「旅行用」と「小さい」)をまとめて1つの記述として扱ってもよい。本発明の観点と記述の抽出規則例では、同じ観点に対して関連する複数の記述をまとめて1つの記述として認定する場合、これらの記述を記号'&&'を用いて例えば'$1&&$2'(ただし、$1，$2は記述)のように示す。

[0026] 次に、観点・記述抽出部120は、上記の観点・記述抽出規則に該当すると認定された観点・記述の組に、観点・記述対が出現したテキストと個々の観点・記述対を識別するための要素メタデータIDを付与し、規則にしたがって抽出する。観点・記述の抽出例を図2(c)の要素メタデータ抽出結果の表に示す。この抽出結果表において、要素メタデータIDの最上段に記載された「1－1a」のうち、左側の「1」は、この観点「開口部」・記述「30cm」がテキスト1から抽出されたものであることを示す。右側の「1a」

の「1」は、観点「開口部」・記述「30cm」がテキスト1を検索したときに第1番目（つまり最初）にヒットした観点・記述であることを示し、「a」は1番目の記述であることを示す。

[0027] なお、本実施例では、要素メタデータIDを＜テキストID＞－＜観点・記述対のテキスト内での番号＞という型式で付与することとしたが、要素メタデータIDの型式は、テキストの識別と観点・記述対の識別が可能なものであれば、これに限定されるものではない。また、統語的属性の付与方法は上記で説明した方法に限定されるものではなく、構文解析や形態素解析を行っても構わない。また、上記の説明は、観点・記述抽出部120が観点・記述抽出規則を用いて文字列の統語的属性を直接判定する例であるが、本発明はこの方法に限定されるものではなく、入力されるテキストにあらかじめ統語的属性を付与しておいてもよいし、属性付与部（後出）で統語的属性を付与してもよい。

[0028] 続いて、メタデータ照合部106は、抽出された要素メタデータの観点間・記述間をそれぞれ比較・照合し、要素メタデータの関連性を推定する。観点・記述の照合方法は、少なくとも観点、記述を構成する文字列の統語的属性を用いて照合するものであれば特に限定されない。例えば、観点または記述の構成語の概念的な類似性をシソーラス、類義語辞書等を用いて比較する方法、また、さらにそれに加えて、観点または記述の構成語の構文的な関係から類似度を推定する方法などを用いることができる。ここでは仮に、観点や記述から助詞や語尾を除く構成語を取り出し、構成語間の構文的関係と、構成語が同義かどうかをメタデータ照合部106内に有するシソーラスを用いて調べた結果を用いて照合することとする。まず、図2(a)のテキスト1、テキスト2の観点から取り出される構成語間の構文的関係は以下のようになる。

[0029] 開口部 →(構成語):開口、部 (構文的関係)連体修飾

ファスナーの開閉 →(構成語):ファスナー、開閉 (構文的関係)連体修飾

皮の感触 →(構成語):皮、感触 (構文的関係)連体修飾

皮の手触り →(構成語):皮、手触り (構文的関係)連体修飾

色合い →(構成語):色合い

[0030] 次に、観点「皮の感触」、「皮の手触り」の構成語のうち「感触」「手触り」をシソーラスにより同義語と認定し、他の構成語「皮」および構文的関係も一致していることから、2

つの観点「皮の感触」「皮の手触り」は同義であり、関連性があると判定することとする。また、記述についても同様にして、同義の記述を求めることとすると、要素メタデータID1－3の「しっとりとやさしい」と要素メタデータID2－2の「しっとりと優しい」という記述が同義であり、関連性があると判定される。なお、要素メタデータの関連性の判定方法は、観点と記述の照合結果に基づいて判定するものならば、上記の方法に限定されるものではなく、他の方法であってもかまわない。例えば、観点や記述の概念的な類似性が数値化されている場合には、観点または記述の数値が一定範囲内にある要素メタデータを「関連性あり」と判定することにしてもよい。

[0031] 次に、要素メタデータ間の関連性に基づいて、メタデータ統合部108が、要素メタデータを統合し、統合メタデータとして統合メタデータ格納部110に格納する。メタデータの統合の仕方は、特に限定されないが、ここでは、

(1)同義の観点をもつメタデータを統一する

(2)同義の観点をもつメタデータで同義の記述があれば統一する

こととする。図2の例では、観点のうち「皮の感触」「皮の手触り」が同義と判定されたので、これらの観点を統合し、例えば「皮の感触」とする。また、これらの観点と対になっている記述「しっとりとやさしい」と「驚くほどなめらかだ」は同義とは見なされないので、統合しない。このようにして統合処理を行なった後の統合メタデータの例を図4に示す。なお、上記の説明では、複数のテキストが入力される場合を説明したが、1テキストが入力されるのであっても構わない。

[0032] このように本実施の形態によれば、テキスト中に表現された事物に関する事実や意見の記述内容を、観点と記述の組として構成し、事実や意見の対応付けや関連性の比較が容易に行える形に整理して抽出し、その抽出結果を用いて、さらに、事実や意見を対応付け、関連する事実や意見を統合することができる。

[0033] (実施の形態2)

図5は本発明の実施の形態2に係る情報抽出システムの構成を示すブロック図である。この情報抽出システム200は、図1に示す実施の形態1に対応する情報抽出システム100と同様の基本的構成を有しており、同一の構成要素には同一の符号を付し、その説明を省略する。

[0034] 本実施の形態の特徴は、入力部102から入力されたテキストの文字列に意味的属性を付与する属性付与部202、前記文字列に意味属性を付与するための意味属性付与規則を記憶した意味属性付与規則格納部204、属性付与部202で付与された意味属性付きテキストを格納する意味属性付きテキスト格納部206を有することである。属性付与部202の処理結果、つまり意味属性が付与されたテキスト(意味属性付きテキスト)は、意味属性付きテキスト格納部206に格納される。この場合、観点・記述抽出部120は意味属性付きテキスト格納部206に格納された意味属性付きテキストに対して観点・記述抽出を行う。

[0035] 属性付与部202は、テキスト中の事物名、数値関連表現(時、数量、金額等)等の文字列を認定し、これらに意味的属性を付与する。事物名や数量表現に意味的属性を付与する方法としては，特に限定されないが、たとえば、キーワード毎にその意味属性を記載した辞書を用いる方法や、文献「福本他:"固有名詞抽出における日本語と英語の比較"、情報処理学会研究会報告98－NL－126, pp. 107－114, 1998」に示される固有名詞抽出技術を利用する方法などを用いることができる。

[0036] ここで意味的属性とは、たとえば、事物名や数量表現を各表現の意味により分類した意味分類である。意味的属性が詳細度のレベルをもつ場合や、該当の表現が一般的な表現の別表現であり、正規化された形を示す必要がある場合は、詳細度レベルや正規化された表現を意味的属性の詳細情報として併記してもよい。

[0037] 以下では、意味属性付与規則を用いて属性付与部202が事物名と数量表現に意味的属性を付与する例を説明する。

[0038] まず、属性付与部202は、意味属性付与規則格納部204に格納された意味属性付与規則を参照して、入力部102から入力されたテキスト内の文字列に対して、規則に該当する意味的属性を持つ表現があるかどうか調べる。その結果、テキスト中の文字列に該当する表現と意味属性をマークし、意味属性付きテキストとして意味属性付きテキスト格納部206に格納する。図6(a)に、入力されたテキスト例、図6(b)に意味的属性を付与されたテキストの例を示す。また、図7に意味属性付与規則の例と意味属性付与規則の構成要素定義の例を示す。なお、構成要素の定義方法は、構成要素名と文字列あるいは文字列パタンのリストの対応付けが可能であれば特に限定さ

れない。例えば、構成要素名と対応する文字列あるいは文字列パタンのリストを1ファイルに記述してもよいし、対応する文字列あるいは文字列パタンのリストは別の複数ファイルに記述しても構わない。なお、これ以降の規則例で同様の構成要素を用いる場合は定義を省略する。

[0039] 図7の意味属性付与規則例には、テキスト中の文字列で該当する意味属性をもつ表現を検出するためのパタン、各パタンに合致する表現の対象部分に付与される意味属性の意味分類および詳細情報が示されている。規則パタンには、意味属性を付与する文字列の文字列表記が、「数字連続」などの文字列パタン、または「製品分類名」などの語リストに対応する、あらかじめ定義した構成要素名が指定されている。なお、規則パタンおよび対象部分の＄1、＄2等の記法は図3の規則と同様である。この例では、詳細情報のうち「val」は数値表現の正規化された値を示し、「unit」は数量単位の表現の正規化形であり、「type」は意味的属性の下位分類を示すこととする。

[0040] 図7の規則を図6(a)のテキスト1に適用した場合、規則1により「20リットル」の意味属性のうち意味分類はQUANT(数量)、詳細情報は[unit=l(単位は'l'の意), val=20(数値は'20'の意)]と認識される。また、規則2により「容量」の意味属性のうち意味分類がQUANT_TYPE(数量分類)として認識される。また、規則3により「A社」の意味属性のうち意味分類がORGANIZATION(組織名)、詳細情報は[type=company(タイプは'会社名'の意)]、等と認識される。認識された結果は、各々該当する意味属性の意味分類と詳細情報を付与され、図6(b)に示したような意味属性付きテキストとして意味属性付きテキスト格納部206に格納される。

[0041] なお、意味属性付与規則の記法は、上記の記法に限定されるものではなく、他の記法であっても構わない。また、意味属性付与規則のパタンの記述方法として、上記説明では文字列パタンや語リストに対応する構成要素名を用いたが、他の記述方法を用いても構わない。また、意味属性付与規則を適用する条件の指定方法として、上記説明ではパタンのみを用いたが、本発明はこれに限定されるものではなく、他の方法であっても構わない。例えば、パタンに加えて、パタンの一部に関する制約を別途指定することとしてもよいし、パタン以外の指定方法を用いてもよい。また、予め意

味属性が付与されたテキストを観点・記述抽出部120に直接入力してもよい。

[0042] 次に、観点・記述抽出部120は、意味属性付きテキスト格納部206に格納された意味属性付きテキストから観点・記述の組を、意味的属性とともに、要素メタデータとして抽出する。意味属性付きテキストの例を図8(a)に、観点・記述認定例を図8(b)に示す。また、観点・記述の抽出のための観点・記述抽出規則の例と観点・記述抽出規則の構成要素の定義例を図9に示す。規則の記法、構成要素の定義方法については図3と同様であり、説明を省略する。

[0043] 図9に示した観点・記述抽出規則と実施の形態1の図3に示した観点・記述抽出規則の違いは、図9では、テキストに付与された意味属性がパタンの一部として記述されていることである。例えば、図9の規則1では＜QUANT__TYPE＞，＜／QUANT__TYPE＞で囲まれたタグ開始記号以外の任意文字列、即ち、QUANT__TYPE(数量分類)という意味属性を付与された文字列が観点として指定される。また、＜QUANT＞，＜／QUANT＞で囲まれたタグ開始記号以外の任意文字列、即ち、QUANT(数量)という意味属性を付与された文字列が前記観点に対応する1つ目の記述として指定されている。図9の規則1を図8(a)のテキスト1に適用した場合、QUANT__TYPEの意味属性を付与された「容量」が観点に相当し、QUANTの意味属性を付与された「20リットル」がこの観点に対応する1つめの記述に相当し、「大きい」が2つ目の記述に相当する。次に、図9の規則3を図8(a)のテキスト1に適用した場合、ORGANIZATIONの意味的属性を付与された文字列「A社」が記述に相当する。この記述に対応する観点はテキスト中には表現されていないが、図9の規則3にしたがって、意味的属性の別名を観点と認定すると、「会社名」が観点と認定される。同様にして、図8(a)の意味属性付きテキスト1，2に対して観点・記述抽出部120が図9の規則を適用して観点と記述を、それらの意味属性である意味分類および詳細情報とともに、識別情報である要素メタデータIDを付与して要素メタデータとして抽出した結果の例を図10に示す。

[0044] なお、上記説明では属性付与部202が文字列の意味的属性を付与するものとしたが、本発明はこれに限定されるものではない。属性付与部202が統語的属性と意味的属性の少なくとも一方をテキストに付与してもよいし、観点・記述抽出部120が観点

・記述抽出規則あるいは他の規則を用いて統語的属性と意味的属性の少なくとも一方を付与してもよいし、入力されるテキストに統語的属性と意味的属性の少なくとも一方があらかじめ付与されていてもよい。

[0045] また、上記説明では意味的属性として意味分類と詳細情報を付与することとしたが、付与される意味的属性は意味分類を含むものであれば、これに限定されるものではなく、例えば詳細情報以外のその他の意味的情報を付与してもかまわない。

[0046] 次に、メタデータ照合部106は、抽出された要素メタデータの観点間・記述間をそれぞれ比較・照合し、関連性を推定する。本実施の形態におけるメタデータ照合部106の照合方法と実施の形態1との違いは、照合の際に要素メタデータの観点や記述の意味属性を用いる点である。ここでは、図10の要素メタデータの観点間、記述間を照合して同義の観点や記述を求める際、実施の形態1の方法に加えて、さらに以下の条件を満たす場合も同義の観点または記述として認定することとする。

・意味分類が「製品名」の表現で、表現中の英数字の境界に「－」が挿入されているかどうかのみが異なるもの。

[0047] 以上の方法により、図10の要素メタデータの観点または記述では、1－2と2－1の観点「製品分類」と記述「バッグ」、1－3の観点「製品名」と記述「A200」と2－2の観点「製品名」と記述「A－200」が各々同義で関連性のある観点と記述と判定され、1－4aと1－4bと2－3の観点「容量」が同義の観点で関連性があると判定される。

[0048] なお、メタデータの観点と記述の照合方法、および要素メタデータ関連性の判定方法は上記の方法に限定されるものではない。観点と記述の照合方法は、例えば、観点または記述の構成語の概念的な類似性をシソーラス、類義語辞書等を用いて比較照合する方法や観点または記述の構成語の構文的な関係から類似度を推定する方法などを用いてもよい。また、要素メタデータの関連性の判定方法は、上記の方法に限定されるものではなく、例えば、観点や記述の概念的な類似性が数値化されている場合には、観点または記述の数値が一定範囲内にある要素メタデータを「関連性あり」と判定することにしてもよい。

[0049] 次に、メタデータ統合部108は、前記要素メタデータの関連性に基づいて、実施の形態1と同様にして、要素メタデータを統合し、統合メタデータとしてメタデータ格納

部110に格納する。ここでは、仮に実施の形態1と同様の条件を満たす観点や記述を統合することとし、詳細な説明は省略する。図10の要素メタデータのうち、関連する要素メタデータを統合してメタデータ格納部110に格納された統合メタデータの例を図11に示す。図11において、同義の観点と記述である1－2と2－1の観点「製品分類」と記述「バッグ」、1－3の観点「製品名」と記述「A200」と2－2の観点「製品名」と記述「A－200」が各々統合されている。また、3つの異なる記述である1－4aの「20リットル」、1－4bの「大きい」、2－3の「不十分だ」の観点「容量」が統合されており、数量である「20リットル」がこの製品の容量として「大きい」「不十分だ」と表現され、図8のテキスト1とテキスト2では異なる評価を受けていることがわかる。

[0050] このように本実施の形態によれば、意味属性付きのテキスト中の文字列に表現された事物に関する事実や意見の記述内容を、観点と記述の意味属性とともに容易に抽出することができる。また、その抽出結果を用いて、関連性をより詳細に判定した上で、関連する事実や意見を統合することにより、事実や意見の対応付けや関連性の比較が容易にできる。

[0051] (実施の形態3)

図12は本発明の実施の形態3に係る情報抽出システムの構成を示すブロック図である。この情報抽出システム300は、図5に示す実施の形態2に対応する情報抽出システム200と同様の基本的構成を有しており、同一の構成要素には同一の符号を付し、その説明を省略する。

[0052] 本実施の形態の特徴は、ユーザからの要求を処理するユーザ要求処理部302と、メタデータを整理してメタデータの出力形式を生成するメタデータ出力形式生成部304と、メタデータ出力形式生成部304が生成したメタデータの出力形式をユーザに提示するメタデータ出力部306と、観点・記述抽出部120の抽出した要素メタデータの話題の事物を推定する話題事物推定部310と話題の事物を推定するための規則である話題事物推定規則を格納した話題事物推定規則格納部312を有することである。

[0053] ここで、「話題事物」とは、各要素メタデータがどの事物について記述されているかという、要素メタデータの話題の事物名のことである。この話題事物は、事物名を表す

要素メタデータのいずれかの記述から選択される。話題事物の候補となりうる事物名は、特に限定されないが、人名、地名、組織名、イベント名、生物や人工物の名およびそれらの分類(例:製品名、製品分類)等がある。

[0054] 上記構成を有する情報抽出システム300について、具体例を用いてより詳細に説明する。今、以下のテキスト1, 2があるとする。

テキスト1:「バッグA200は容量が不十分だし、バッグA300は容量があまりに大きい。」

テキスト2:「バッグA200は容量が20リットルで、バッグA300の容量は30リットル。」

前記テキストが入力部102から入力され、属性付与部202で意味属性が付与され、観点・記述抽出部120で観点・記述が認定され、要素メタデータが抽出されるまでの処理の流れは実施の形態2と同様であり、説明を省略する。図13(a)に上記テキストに対して意味分類を付与し、観点・記述を認定した結果の例と、図13(b)に要素メタデータの抽出結果の例を示す。

[0055] 次に、話題事物推定部310は、話題事物推定規則格納部312に格納された話題事物推定規則にしたがって、テキスト内の話題事物を推定する。話題事物の推定方法は話題事物推定規則を用いるものであれば、特に限定されない。話題事物推定部310が話題事物推定規則を用いて直接話題事物を推定することとしてもよいし、まず話題事物候補となる要素メタデータの種類を決定し、その後に話題事物推定規則を用いて推定することとしてもよい。その場合、入力されるテキストが、例えば会社名と人名等、複数の種類の話題をもつ可能性がある場合は、複数の話題事物推定候補を想定し、前記話題事物推定部310が、適当な話題事物を選択できるようにしておくことが望ましい。例えば、話題事物候補が「観点が製品名または人名」である要素メタデータの記述と規定されている場合、観点が製品名または人名である要素メタデータの記述が話題事物の候補と規定されているとする。この場合、テキスト1, 2とも、製品名を観点にもつ要素メタデータの記述であり、「A200」、「A300」が話題事物候補となる。

[0056] 以下では、話題事物推定部310が、話題事物推定規則格納部312に格納された話題事物推定規則にしたがって、テキスト内の話題事物を推定する場合について説

明する。ここでは仮に、条件部に記述したパタンとのマッチングにより話題事物推定を行うこととし、図14(a)に話題事物推定規則と図14(b)に話題事物推定規則構成要素定義の例を示す。なお、規則の条件部のパタンの記法や構成要素の定義方法は図3と基本的に同様であるが、図14の規則2、規則3では条件としてパタンのみでなく、パタンの一部文字列が同一であることも条件に加えた。

図14(a)の規則を用いて図13(a)のテキスト1，2から、図13(b)の要素メタデータの話題事物を推定する。例えば、テキスト1に図14の規則1を適用すると、まず2番目の記述である＜DESC2＞＜PROD_NAME＞A200＜／PROD_NAME＞＜／DESC2＞が規則1の条件部に記述されたパタンに合致し、同規則にしたがって、このうち、＄1に相当する「A200」の話題事物は「A200」自体と推定される。同様にして、図14の規則を用いて、図13(a)のテキスト1，2から、図13(b)の要素メタデータの話題を推定した例を図15に示す。図15の要素メタデータID1－1、1－4、2－1、2－4の要素メタデータについては図14の規則3が適用され、図15のID1－2、1－5、2－2、2－5の要素メタデータについては図14の規則1が適用され、図15のID1－3、1－6、2－3、2－6のメタデータについては図14(a)の規則2が適用されている。

[0057] なお、話題事物の推定方法は、話題事物抽出規則を用いるものならば上記に限定されるものではなく、例えば要素メタデータの観点・記述や統語的属性や意味的属性あるいは他の属性を上記とは別の記法の規則に指定しても構わない。また、話題事物候補の種類によって異なる規則を適用することとしてもよい。

[0058] 続いてメタデータ照合部106が、抽出された要素メタデータの観点間・記述間のそれぞれを比較・照合し、関連性を推定する。要素メタデータの観点・記述の照合方法は実施の形態1または2と基本的に同様であるが、本実施の形態では、さらに話題事物の推定結果をも用いて照合する。

[0059] 図15の例では、要素メタデータID1－1，1－2，1－3，2－1，2－2，2－3が同じ話題事物「A200」を持ち、1－4、1－5、1－6、2－4、2－5、2－6が同じ話題事物「A300」を持つ。同じ話題事物を持つ要素メタデータ毎に、実施の形態1と同様にして、同義の観点や記述を求めることとすると、まず、話題事物が「A200」である要素メ

タデータについては、同義の観点と記述を持つ要素メタデータは1－1と2－1、1－2と2－2である。また、同義の観点を持つ要素メタデータは、1－3と2－3が得られる。前者の観点と記述、後者の観点は各々関連性があると推定される。

[0060] 同様に、話題事物が「A300」である要素メタデータについては、同義の観点と記述を持つ要素メタデータは1－4と2－4、1－5と2－5である。また、同義の観点を持つ要素メタデータは、1－6と2－6が得られる。前者の観点と記述、後者の観点は各々関連性があると推定される。

[0061] なお、メタデータ照合部106の照合方法および関連性の推定方法は上記に限定されるものではない。上記説明では、同じ話題事物をもつ要素メタデータ毎に、同義の観点や記述を求めたが、例えば、同義の観点や記述を持つ要素メタデータを求めた後で、同じ話題事物を持つものを求めてもよいし、要素メタデータの意味属性等をさらに用いても構わない。

[0062] 次に、実施の形態1と同様にして、メタデータ統合部108が要素メタデータを統合し、統合メタデータとしてメタデータ格納部110に格納する。要素メタデータの統合の仕方は限定されないが、ここでは例として、

(1)同じ話題をもつ事物を統合する、

(2)同じ話題で同義の観点をもつ要素メタデータを統一する、

(3)同じ話題で同義の観点をもつ要素メタデータで同義の記述があれば統一する。

[0063] この例を用いた場合について説明する。図15の要素メタデータのうち、同じ話題事物をもつ1－1、1－2、1－3、2－1、2－2、2－3は上記(1)により話題事物を統合する。同様に、1－4、1－5、1－6、2－4、2－5、2－6も話題事物が統合される。次に、同じ話題事物と同義の観点をもつ要素メタデータ1－1と2－1、1－2と2－2、1－3と2－3、1－4と2－4、1－5と2－5、1－6と2－6は上記(2)にしたがって、各々話題事物と観点が統合される。さらに、同じ話題事物をもち、同義の観点と記述をもつ要素メタデータ1－1と2－1、1－2と2－2、1－4と2－4、1－5と2－5は上記(3)にしたがって、各々話題事物と観点と記述が統合される。

[0064] 以上のようにして、テキスト1，2から抽出された図15の要素メタデータをメタデータ統合部108が統合した結果、メタデータ格納部110に格納された統合メタデータの

例を図16に示す。この統合結果から、「A200」の「容量」が「20リットル」で「不十分だ」と評価されている一方、「A300」の「容量」が「30リットル」で「あまりに大きい」と評価されていることがわかる。なお、メタデータの統合方法は上記に限定されるものではなく、メタデータ照合部106が推定した要素メタデータの観点と記述の関連性に基づいて統合を行うものであれば他の方法であっても構わない。例えば同義の観点や記述をもつ要素メタデータをまず統合し、その後、同じ話題事物をもつ要素メタデータを統合するようにしてもよい。

[0065] 次に、ユーザ要求処理部302は、ユーザ要求が入力され、ユーザの要求した出力形式をメタデータ出力形式生成部304に出力する。メタデータ出力形式生成部304は、メタデータ格納部110に格納された統合メタデータを参照して、ユーザの要求した出力形式でメタデータを生成し、メタデータ出力部306を通じてユーザに提示する。

[0066] ここでは、ユーザ要求の指定にしたがって、メタデータの出力形式の一例としてメタデータテーブルを生成する場合を説明する。まず、ユーザ要求処理部302を通じてユーザ要求が入力される。ユーザ要求処理部302に入力されるユーザ要求は、話題事物を含めた要素メタデータの一部、あるいはこれらの組み合わせのいずれかを指定するものとする。今、ユーザ要求の例として、例えば「(話題事物:A200) かつ (観点:容量)」という条件が要素メタデータの満たすべき条件として指定されたものとする。ユーザ要求処理部302は、指定されたユーザ要求の指定形式をチェックし、問題がなければユーザ要求をメタデータ出力形式生成部304に送る。

[0067] なお、この例では、ユーザ要求が上記の形式で入力されることとしたが、ユーザ要求が自由なテキスト(例:「A200の容量が知りたい」)で入力されても構わない。後者の場合は、ユーザ要求処理部302がテキストを直接解析して上記の条件を取り出すこととしてもよい。また、ユーザ要求処理部302が、入力部102にユーザの入力したユーザ要求のテキストを一旦送り、観点・記述抽出部120によって抽出された要素メタデータと、それらの構文的な関係から指定された条件の内容を解析することとしてもよい。

[0068] メタデータ出力形式生成部304は、ユーザ要求処理部302から受け取ったユーザ

要求の指定内容にしたがって、メタデータ格納部110に格納された統合メタデータの中から該当する要素メタデータを選別し、選別されたメタデータを出力形式に対応させて生成する。例えば、ユーザ要求の内容に話題事物の指定があれば、この話題事物を話題にもつ要素メタデータを統合メタデータの中から選別し、指定された観点や記述の条件を満たす要素メタデータをさらに選別し、それらを対象としたメタデータテーブルを生成する。メタデータ出力部306が生成されたメタデータテーブルを出力する。

[0069] 図17に図16の統合メタデータのうち、ユーザ要求(話題事物:A200)かつ(観点:容量)を満たす要素メタデータのみを取り出して作成したメタデータテーブルの例を示す。この場合は、話題事物が「A200」で観点が「容量」の要素メタデータのみがテーブルとして出力されている。なお、上記の説明ではメタデータの出力形式はメタデータのテーブルとして説明したが、出力形式はテーブル以外の他の形式であっても構わない。

[0070] このように本実施の形態によれば、テキスト中に表現された事物に関する事実や意見の記述内容を、推定された話題の事物とともに、事実や意見の対応付けを容易に行うことができる。また、その抽出結果を用いて、さらに、事実や意見を話題事物毎により精密に対応付け、関連性をより詳細に判定した上で、関連する事実や意見を統合することができ、関連性の比較が容易に行える形に整理して抽出することができる。

また、話題事物を含めた要素メタデータを、ユーザの指定にもとづいて整理したメタデータ出力形式をユーザに提示することにより、ユーザが要求する情報を整理して提示することができる。

[0071] (実施の形態4)

図18は本発明の実施の形態4に係る情報抽出システムの構成を示すブロック図である。この情報抽出システム400は、図12に示す実施の形態3に対応する情報抽出システム300と同様の基本的構成を有しており、同一の構成要素には同一の符号を付し、その説明を省略する。

[0072] 本実施の形態の特徴は、前記入力部102がソース情報およびユーザ情報をも受け

とり、メタデータ照合部106が要素メタデータ、ソース情報またはユーザ情報を用いて観点・記述の客観性と信頼性を判定する客観性・信頼性判定部412と、客観性と信頼性を評価するための客観性・信頼性判定規則を格納した客観性・信頼性判定規則格納部414を有することである。

[0073] ここで、ソース情報とは、入力されるテキストに関する書誌事項の情報を指すものとし、テキスト中のソース情報の記述をソース情報記述と呼ぶこととする。ソース情報の例としては、テキストの種別、入手元、作成者分類、作成者、組織名、作成日時等がある。ソース情報記述は、テキストとの対応付けが可能な形であれば、入力テキストの一部として入力されてもよいし、入力テキストとは別に入力されてもよい。ソース情報記述の書式は、特に限定されないが、テキストの識別情報とともに入力されることとする。

[0074] また、ユーザ情報とは、入力されるテキストの作者に関する情報を指すものとし、テキスト中に表現されたユーザ情報の記述をユーザ情報記述と呼ぶこととする。ユーザ情報の例としては、ユーザの性別、年齢、職業、勤務先、趣味、等がある。ユーザ情報記述はテキストとの対応付けが可能な形であれば、テキストの一部として入力されてもよいし、入力テキストとは別に入力されてもよい。ユーザ情報記述の書式は、特に限定されないが、テキストの識別情報とともに入力されることとする。

[0075] また、ソース情報記述、ユーザ情報記述を観点と記述の組として構成したものを各々ソースメタデータ、ユーザメタデータと呼ぶ。ソースメタデータおよびユーザメタデータには、対応するテキストと個々のソースメタデータまたはユーザメタデータを識別するためのソースメタデータID、またはユーザメタデータIDが付与される。ソースメタデータIDおよびユーザメタデータIDの書式は特に限定されないが、テキストとの対応関係をとる必要があるため、対応するテキストIDが推定可能な書式とすることが望ましい。

[0076] 客観性・信頼性判定部412は、要素メタデータ、ソースメタデータ、またはユーザメタデータのいずれかを用いて要素メタデータの観点・記述の客観性と信頼性を判定し、判定結果を要素メタデータの評価データとする。メタデータ統合部108は、要素メタデータに加えて、ソースメタデータ、ユーザメタデータ及び要素メタデータの評価デ

ータをも統合メタデータの結果に含めることができる。また、ユーザは、ユーザ要求処理部302から要素メタデータだけでなく、ユーザメタデータやソースメタデータや要素メタデータの評価データをも用いて必要な情報を指定し、対応するメタデータ出力形式を得ることができる。

[0077] 次いで、上記構成を有する情報抽出システム400について、具体例を用いてより詳細に説明する。本実施の形態においては、ソース情報記述およびユーザ情報記述が入力テキストの一部として入力されることとし、入力テキストの特定のブロックにソース情報記述、ユーザ情報記述が各々記述されていることとする。

[0078] 入力部102からテキストが入力される。

属性付与部202は、入力されたソース情報記述、またはユーザ情報記述を含むテキストに意味属性付与規則を用いて意味属性を付与し、意味属性付きテキストを意味属性付きテキスト格納部206に出力する。図24(a)にソース情報記述およびユーザ情報記述のブロックを除いたテキスト例1～4を示す。ここまでの処理の流れは、実施の形態2乃至3と同様であるので、詳細な説明を省略する。図19(a)にソース情報記述、図19(b)にユーザ情報記述の例、図19(c)に意味属性付きソース情報記述の例、図19(d)に意味属性付きユーザ情報記述の例をそれぞれ示す。また、図20(a)にソース意味属性付与規則、図20(b)にユーザ意味属性付与規則の例を示す。

[0079] 次に、観点・記述抽出部120が、観点・記述抽出規則格納部122に格納された観点・記述抽出規則、ソース観点・記述抽出規則、またはユーザ観点・記述抽出規則を用いて、意味属性付きテキスト格納部206に格納された意味属性付きテキスト、意味属性付きソース情報または意味属性付きユーザ情報から、少なくとも観点と記述の組からなる要素メタデータ、ソースメタデータ、またはユーザメタデータをそれぞれ抽出する。

[0080] まず、前記意味属性付きテキストのソース情報記述およびユーザ情報記述のブロックから観点・記述抽出部120がソースメタデータとユーザメタデータを抽出する場合について説明する。ソースメタデータまたはユーザメタデータを抽出する際、図22(a)に示すように各ソースメタデータにはソースメタデータID、図22(b)に示すようにユーザメタデータにはユーザメタデータIDが付与される。なお、本実施の形態では、ソ

ースメタデータIDおよびユーザメタデータIDを、各々、＜テキストID＞－S＜観点・記述対のソース情報内での番号＞、＜テキストID＞－U＜観点・記述対のユーザ情報内での番号＞という形式で付与することとしたが、ソースメタデータIDの形式およびユーザメタデータIDの形式は、これに限定されるものではない。

[0081] 図21(a)にソース観点・記述抽出規則および図21(b)にユーザ観点・記述抽出規則の例を示す。図21(a)のソース観点・記述抽出規則および図21(b)のユーザ観点・記述抽出規則では、実施の形態1の観点・記述抽出規則と同様に、規則のパタンには、観点・記述に相当する文字列またはその周辺の文字列の統語的属性および意味的属性が指定されている。なお、文字列の統語的属性の指定方法として、図21(a)(b)では表記が用いられ、意味的属性の指定方法としては意味属性の意味分類と詳細情報が用いられているが、本発明はこれに限定されるものではなく、統語的属性と意味的属性のどちらか一方のみを指定しても構わないし、例えば統語的属性として品詞分類等を用いても構わない。

[0082] 以下、図19(c)の意味属性付きソース情報記述および図19(d)の意味属性付きユーザ情報記述から図21(a)のソース観点・記述抽出規則または図21(b)のユーザ観点・記述抽出規則を用いてソースメタデータおよびユーザメタデータを抽出する場合について説明する。例えば、図19(c)の意味属性付きソース情報記述に図19(a)のソース観点・記述抽出規則1を適用する場合、図19(c)の文字列<URL type=会社webページs>http://www.aaa.co.jp/article1</URL>が前記規則1のパタンに該当し、そのうち、パタン中の最初の'()'で括られた部分に相当するhttp://www.aaa.co.jp/article1が、規則中で指定された観点「テキストの入手元」に対応する記述に相当する。

[0083] 図19(c)の意味属性付きソース情報記述および図19(d)の意味属性付きユーザ情報記述から図21(a)のソース観点・記述抽出規則または図21(b)のユーザ観点・記述抽出規則を用いて抽出したソースメタデータ抽出結果とユーザメタデータ抽出結果の例を各々、図22(a)、図22(b)に示す。

[0084] 次に、前記意味属性付きテキストのソース情報記述、ユーザ情報記述のブロック以外から観点・記述抽出部120が要素メタデータを抽出し、話題事物推定部310が話

題事物を推定するまでの流れについて説明する。図24(b)に図24(a)の各テキストに属性付与部202が意味属性を付与した例、図25に観点・記述抽出規則の例を示す。図24(b)の意味属性付きテキストから、図25の観点・記述抽出規則を用いて、実施の形態2または3と同様にして観点・記述を抽出する。例えば図24(b)の意味属性付きテキスト1に図25の規則1を適用すると、観点「容量」に対して、2つの記述「20リットル」「大きい」が抽出される。同様にして図24(b)の意味属性付きテキスト1～4から図25の規則により図26に示すような観点と記述が抽出される。さらに、図24(b)の意味属性付きテキスト1～4から図14の話題事物推定規則を用いて実施の形態3と同様にして推定する。

[0085] 図24(b)の意味属性付きテキスト1～4から、観点・記述抽出部120が抽出した観点・記述、およびそれらの意味的属性と、話題事物推定部310が推定した話題事物をまとめて要素メタデータの例として図26に示す。なお、図26では、要素メタデータの一部のみを示している。また、実施の形態3では、話題事物の推定に関して、テキストから得られる情報だけを用いて推定する方法を説明したが、他にソース情報やユーザ情報から得られるメタデータを用いてもよい。

[0086] 次に、メタデータ照合部106の客観性・信頼性判定部412は、観点・記述抽出部120において意味属性付きテキストから抽出された要素メタデータとソースメタデータとユーザメタデータのうち、少なくとも1つを用いて、客観性・信頼性判定規則格納部414に格納された客観性・信頼性判定規則にしたがって前記要素メタデータの客観性および信頼性を判定する。

[0087] ここで、要素メタデータの客観性とは、要素メタデータが客観的に記述されているかどうかを示し、例えば事実として記述されているならば客観性は高く、意見として記述されているならば客観性は低いと考えられる。客観性は、数値として表現してもよいし、ある閾値または判定条件により「事実」「意見」等の分類で表現してもよい。

[0088] また、要素メタデータの信頼性とは、要素メタデータが信頼できるかどうかを示し、例えば個人のホームページに意見として書かれた記述の信頼性は比較的低く、新聞記事に事実として書かれた記述の信頼性は高い、等と考えられる。なお、信頼性は、数値として表現してもよいし、ある閾値または判定条件により「信頼性高」「信頼性低」等

の分類で表現してもよい。

[0089] 要素メタデータの客観性・信頼性の判定には、少なくとも要素メタデータ、ソースメタデータ、ユーザメタデータのいずれか一つを用いることとするが、これらの他に文字列の統語的属性、意味的属性や統計的な情報等を組み合わせて用いても構わない。

[0090] 図23に客観性・信頼性判定規則の例を示す。ここでは、客観性を1～0(1は客観性が高く、0は低いものとする)、信頼性を1～0(1は信頼性が高く、0は低いものとする)で表現する。例えば、規則4は観点が「用途」で記述の意味分類が「USAGE」であるような要素メタデータについて、ソースメタデータのテキストの入手元が「会社webページ」であれば客観性は1、信頼性も1と判定する規則である。

[0091] 次に、図26の要素メタデータについて、テキストの要素メタデータ、ソースメタデータ、統語的属性により、客観性・信頼性判定規則を用いて、客観性・信頼性の判定処理を行った例を説明する。

[0092] 今、観点・記述抽出部120により、入力テキストのうち、図26の要素メタデータの抽出元のテキスト1～4に対応するソース情報記述およびユーザ情報記述のブロックから、それぞれ以下のようなソースメタデータとユーザメタデータが抽出されているとする。

[0093] テキスト1

ソースメタデータ

観点: テキストの入手元

記述の意味属性:会社webページ

テキスト2

ソースメタデータ

観点:テキストの入手元

記述の意味属性:個人webページ

ユーザメタデータ

観点:性別

記述:男性

テキスト3

ソースメタデータ

観点:テキストの入手元

記述の意味属性:個人webページ

ユーザメタデータ

観点:性別

記述:女性

テキスト4

ソースメタデータ

観点:テキストの入手元

記述の意味属性:個人webページ

ユーザメタデータ

観点:性別

記述:男性

[0094] 上記のソースメタデータおよびユーザメタデータを用いて、図26の要素メタデータの客観性および信頼性を、図23の客観性・信頼性判定規則を用いて判定する。例えば、図26の要素メタデータIDが1－3aの要素メタデータの場合、要素メタデータの観点が「容量」、記述の意味分類が「QUANT」であり、抽出元のテキスト1は会社webページであるので、図23の規則6が適用され、客観性、信頼性、ともに1と判定される。一方、図26の要素メタデータIDが1－3bの要素メタデータの場合、要素メタデータの観点が「容量」、記述の意味分類が「なし」であり、抽出元のテキスト1は会社webページであり、さらに要素メタデータを含む文の「文末が不確定表現1以外」であるので、図23の規則9が適用され、客観性は0、信頼性は0．5と判定される。同様にして、上記のソースメタデータおよびユーザメタデータを用いて、図26の要素メタデータに対して、図23の客観性・信頼性判定規則を用いて客観性・信頼性判定部412が判定した客観性・信頼性判定結果例を図27に示す。なお、規則の記法や構成要素定義については図3、図7等と同様であり、説明を省略する。

[0095] また、客観性・信頼性判定規則の条件として、上記説明ではテキストの要素メタデ

ータとソースメタデータと統語的属性を用いたが、要素メタデータとソースメタデータとユーザメタデータの少なくとも1つを含むものであれば、本発明はこれに限定されない。また、図23の客観性・信頼性判定規則では、ソースメタデータの観点「テキストの入手元」と対応する記述の意味属性を規則の条件の一部に用いたが、他の観点と記述の組を用いてもよい。例えば「作成日」を用いて作成日が古い要素メタデータは信頼性が低いと判定する、あるいは「作成者」を用いて特定の人の書いたテキストの信頼度を上げる、または下げるというようにしてもよい。また、要素メタデータと他の情報を組み合わせる場合、例えば、統計的な情報と組み合わせて、同じ観点に対して多数の類似の内容の記述をもつ要素メタデータの信頼度を上げる。あるいは、多数の人の記述と異なる内容の記述をもつ要素メタデータの信頼度を下げるようにしてもよい。なお、図23の客観性・信頼性判定規則では、1規則で客観性と信頼性を同時に判定しているが、客観性の判定規則と信頼性の判定規則を分けて、1規則でいずれか一方を判定するようにしても構わない。

[0096] 次にメタデータ照合部106は、抽出された要素メタデータの観点間・記述間をそれぞれ比較・照合し、関連性を推定する。メタデータ照合部106の観点・記述の照合方法は特に限定されない。ここでは実施の形態1または2または3と同様とするが、客観性・信頼性のデータをも用いて、観点・記述間の照合結果から関連性が高いと推定される要素メタデータのうち、客観性や信頼性の値が近いものはさらに関連性が高い、と推定するようにしてもよい。

[0097] また、上記の説明ではソースメタデータとユーザメタデータは、客観性や信頼性の判定のみに用いたが、これらをメタデータ照合部106が要素メタデータの比較・照合を行う際に直接用いてもよい。例えば、複数の個人webページから抽出された要素メタデータのある製品の容量についての記述がある場合、ユーザメタデータの「性別」の記述が同じであったり、「年齢」の記述が一定範囲にあれば関連性が高い、というようにしてもよい。

[0098] 次にメタデータ統合部108は、要素メタデータとソースメタデータとユーザメタデータと評価を含めた要素メタデータを統合し、統合結果を統合メタデータ格納部110に格納する。

[0099] 統合の仕方は特に限定されないが、ここでは例として、以下の(1)～(4)とする。

(要素メタデータ)

(1)同じ話題をもつメタデータを統合する

(2)同じ話題で同義の観点をもつメタデータを統合する

(3)同じ話題で同義の観点をもつメタデータで同義の記述があれば統合する

(4)同じ話題で同義の観点と同義の記述をもつメタデータで意味属性が同じなら統合する

[0100] 統合の仕方を(1)～(4)とした場合に、図27の要素メタデータをメタデータ統合部108が統合する場合について説明する。まず、図27のメタデータは、すべて同じ話題「A200」を持つので、上記(1)により、共通の話題で統合される。次に、同じ話題をもつ各要素メタデータの観点が同義であるかどうかを実施の形態1と同様にして判定する。図27の例では、観点は「製品分類」、「製品名」、「容量」、「用途」の4種類のみであり、これらは同義ではないので、これら4つの観点をもつ要素メタデータを各々統合すると、観点「製品分類」で要素メタデータ1－1、2－1、3－1、4－1が統合され、観点「製品名」で要素メタデータ1－2、2－2、3－2、4－2が統合され、観点「容量」で要素メタデータ1－3a、1－3b、2－3、3－3、4－3が統合される。

[0101] 次に、同じ話題で同義の観点をもつメタデータの記述が同義であるかどうかを実施の形態1と同様にして判定する。図27の例では、例えば、話題「A200」で同義の観点「製品分類」をもつ要素メタデータの記述はすべて「バッグ」であるのでこれらは同義とみなされ、上記(3)により、要素メタデータ1－1、2－1、3－1、4－1の記述は統合される。同様に、話題「A200」で同義の観点「製品名」をもつ要素メタデータ1－2、2－2、3－2、4－2の記述、および同義の観点「用途」をもつ要素メタデータ3－4、4－4も各々統合される。一方、例えば、話題「A200」で同義の観点「容量」をもつ要素メタデータの記述「20リットル」、「大きい」、「海外出張用―不十分だ」、「国内出張用―あまりに大きい」、「国内出張用―十分だ」は同義と判定されないので、統合されない。

[0102] 次に、話題「A200」で同義の観点「製品分類」と同義の記述「バッグ」をもつ要素メタデータの意味分類はすべて「PROD_TYPE」であるのでこれらは同義とみなされ

、上記(4)により、要素メタデータ1－1、2－1、3－1、4－1の意味分類は統合される。同様に、話題「A200」で同義の観点「製品名」と同義の記述「A200」をもつ要素メタデータ1－2、2－2、3－2、4－2の意味分類、および同義の観点「用途」と同義の記述「国内出張用」をもつ要素メタデータ3－4、4－4の意味分類も各々統合される。

[0103] 以上のようにして、図27のメタデータをメタデータ統合部108が統合した結果、統合メタデータ格納部110に格納されたメタデータの統合結果の例を図28に示す。なお、図28において、要素メタデータの一部は記述を省略している。

[0104] 図28の例では、客観性、信頼性ともに高い情報、すなわち事実である可能性が高い情報として、「A200」という「バッグ」の「容量」が「20リットル」であるという情報がある。また、それに関する客観性の低い情報、すなわち意見と思われる情報として、会社のホームページではその容量が「大きい」と評価されているが、個人のホームページでは、「海外出張用」には男性1名に「不十分だ」と評価され、「国内出張用」には女性1名に「あまりに大きい」と評価され、男性1名に「十分だ」と評価されていることが分かる。

[0105] 次に、メタデータ出力形式生成部304は、ユーザ要求処理部302からユーザ要求の指定があればユーザ要求の指定にしたがってメタデータ出力形式を生成し、メタデータ出力部306を通じてユーザに提示するが、ここまでの流れは実施の形態3と同様である。ただし、本実施の形態では、要素メタデータの評価データをもユーザ要求として指定することができる。ここでは、図28のメタデータの統合結果から、次のような要素メタデータの評価データを含むユーザ要求の指定を受けて、メタデータ出力形式生成部304がユーザの指定した条件に合致するメタデータについてメタデータテーブルを生成する場合を一例として説明する。

[0106] ユーザ要求の指定

話題事物:A200

客観性:0

テキストの種類:個人webページ

この指定は、個人webページのテキストに書かれた、「A200」という事物についての

評価データとしては「客観性が0」の記述、すなわち意見を求めるものである。なお、上記はユーザ要求の指定方法の一例であり、指定方法は上記に限定されるものではない。

[0107] 上記のユーザ要求の指定により、実施の形態3と同様にして生成されたメタデータテーブルの例を図29に示す。図29のメタデータテーブルから、A200という事物についての個人webページのテキストに書かれた意見としては、容量と用途という観点が取り上げられていること、用途としては海外出張用、国内出張用という2つの用途について評価されていること、容量は海外出張用には不十分(男性1名)と評価され、国内出張用にはあまりに大きい(女性1名)、十分だ(男性1名)と評価が分かれていることがわかる。

[0108] このように本実施の形態によれば、テキスト中に表現された事物に関する事実や意見の記述内容を、推定された話題の事物とともに、事実や意見を対応付けて抽出することができる。また、抽出された事実や意見の関連性の比較が容易に行える形で抽出し、事実や意見を話題事物毎に対応付けた上で、客観性や信頼性の評価結果をも含めてユーザに提示する。これにより、ユーザが提示された情報を適切に解釈し、またユーザが必要な情報のみを的確に選択することができる。

[0109] 本発明は、図面に示す好ましい実施例に基づいて説明されてきたが、当業者であれば、この発明を容易に変更及び改変し得る事は明らかであり、そのような変更部分も発明の範囲に含まれるものである。

産業上の利用可能性

[0110] 本発明にかかる情報抽出システムは、観点・記述抽出部、観点・記述抽出規則格納部、メタデータ格納部を有し、情報抽出システム、情報検索システムとして有用である。また、情報分析／評価システム、情報配信システム等の用途にも応用できる。

請求の範囲

[1] テキストを入力する入力部と、

テキストに記述された表現の観点とその観点に関する記述の組を特定するための観点・記述抽出規則を格納する観点・記述抽出規則格納部と、

前記入力部に入力されたテキスト中の文字列に付与された統語的属性または意味的属性の少なくとも一方の属性から、前記観点・記述抽出規則を用いて観点とその記述の組を対応付けた要素メタデータとして抽出する観点・記述抽出部と、

前記観点・記述抽出部が抽出した要素メタデータを格納するメタデータ格納部と

を具備する情報抽出システム。

[2] 前記統語的属性は少なくとも文字列表記または品詞分類のいずれかを含む請求項1記載の情報抽出システム。

[3] 前記意味的属性は少なくとも意味分類を含む請求項1記載の情報抽出システム。

[4] 前記観点・記述抽出部は、対応付けた観点と記述の組を要素メタデータとして抽出する際に、対応付けた観点と記述の組を識別するための識別情報(要素メタデータID)を付与して抽出する請求項1記載の情報抽出システム。

[5] テキストから任意の文字列を抽出し、少なくとも文字列の意味分類を特定するための意味属性付与規則を用いて、文字列毎に意味属性を付与した意味的属性付きテキストを出力する属性付与部をさらに有する請求項1記載の情報抽出システム。

[6] 前記観点・記述抽出部は、テキスト中に観点が表現されず、記述のみが表現されている場合に、記述の意味的属性を観点として、観点と記述の組を抽出する請求項1記載の情報抽出システム。

[7] 前記観点・記述抽出部の抽出した要素メタデータの観点間と記述間をそれぞれ照合し、関連性を推定するメタデータ照合部と、

前記推定された関連性に基づいて、関連性のある要素メタデータを統合し、統合メタデータを出力するメタデータ統合部をさらに具備する請求項1に記載の情報抽出システム。

[8] 前記メタデータ照合部は、前記観点・記述抽出部の抽出した要素メタデータの観点と記述を照合する際に、少なくとも観点、記述を構成する文字列の意味的属性を用い

て照合を行う請求項6記載の情報抽出システム。

[9] 前記観点・記述抽出部で抽出された要素メタデータに対して、話題の事物を推定するための話題事物推定規則を用いて、話題とされた事物を推定する話題事物推定部、をさらに有し、

前記メタデータ格納部が要素メタデータとともに、前記話題事物推定部で推定された話題の事物をも対応付けて格納する請求項7に記載の情報抽出システム。

[10] 前記話題事物推定部は、要素メタデータの話題の事物を推定する際に、

前記メタデータ格納部に格納された要素メタデータの観点・記述や意味的属性から、話題の事物を推定する請求項9に記載の情報抽出システム。

[11] 前記メタデータ照合部が、観点と記述の照合を行う際に、前記話題事物推定部によって推定された話題事物単位で前記観点と記述の照合を行う請求項9または10に記載の情報抽出システム。

[12] 前記観点・記述抽出規則が、テキストの作者に関する情報であるユーザ情報を特定するための規則であるユーザ観点・記述抽出規則を含み、

前記観点・記述抽出部が前記ユーザ観点・記述抽出規則を用いて、ユーザ情報に関する要素メタデータであるユーザメタデータを抽出する、請求項1記載の情報抽出システム。

[13] 前記観点・記述抽出規則が、テキストの書誌事項に関する情報であるソース情報を特定するための規則であるソース観点・記述抽出規則を含み、

観点・記述抽出部が前記ソース観点・記述抽出規則を用いて、ソース情報に関する要素メタデータであるソースメタデータを抽出する、請求項1記載の情報抽出システム。

[14] 前記メタデータ照合部が、少なくとも要素メタデータ、またはユーザメタデータ、またはソースメタデータのうち1種類以上を用いて、観点、記述の客観性と信頼性を判定する客観性・信頼性判定部と、前記観点、記述の客観性・信頼性を判定するための客観性・信頼性判定規則を格納する客観性・信頼性判定規則格納部をさらに有する請求項9に記載の情報抽出システム。

[15] 前記話題事物推定部が要素メタデータの話題の事物を推定する際に、

前記メタデータ格納部に格納された要素メタデータに加えて、ソースメタデータまたはユーザメタデータの少なくともいずれか一方を用いて話題の事物を推定する請求項9に記載の情報抽出システム。

[16] 前記メタデータを表形式に整理してメタデータテーブルを生成するメタデータ出力形式生成部と、生成されたメタデータテーブルをユーザに提示するメタデータ出力部と、をさらに有する請求項14に記載の情報抽出システム。

[17] ユーザからの要求を処理するユーザ要求処理部をさらに有し、前記メタデータ出力形式生成部が、前記ユーザ要求処理部を通じて入力されたユーザ要求に合致するメタデータを用いてメタデータテーブルを生成する請求項16に記載の情報抽出システム。

[18] テキストを入力するステップと、

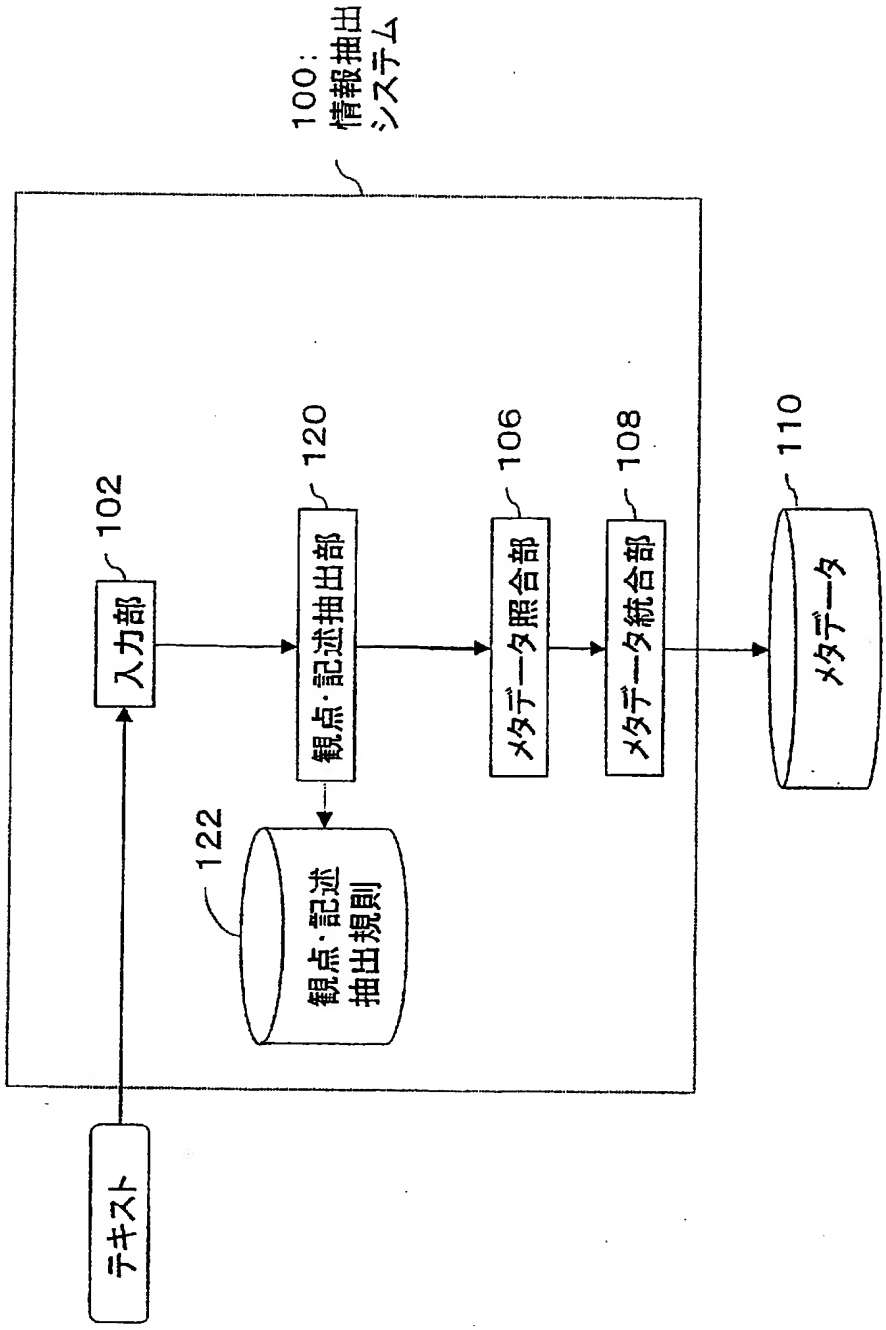
テキストに記述された表現の観点とその観点に関する記述の組を特定するための観点・記述抽出規則を参照するステップと、

前記入力部に入力されたテキスト中の文字列に付与された統語的属性または意味的属性の少なくとも一方の属性から、前記観点・記述抽出規則を用いて観点とその記述の組を対応付けた要素メタデータとして抽出するステップと、

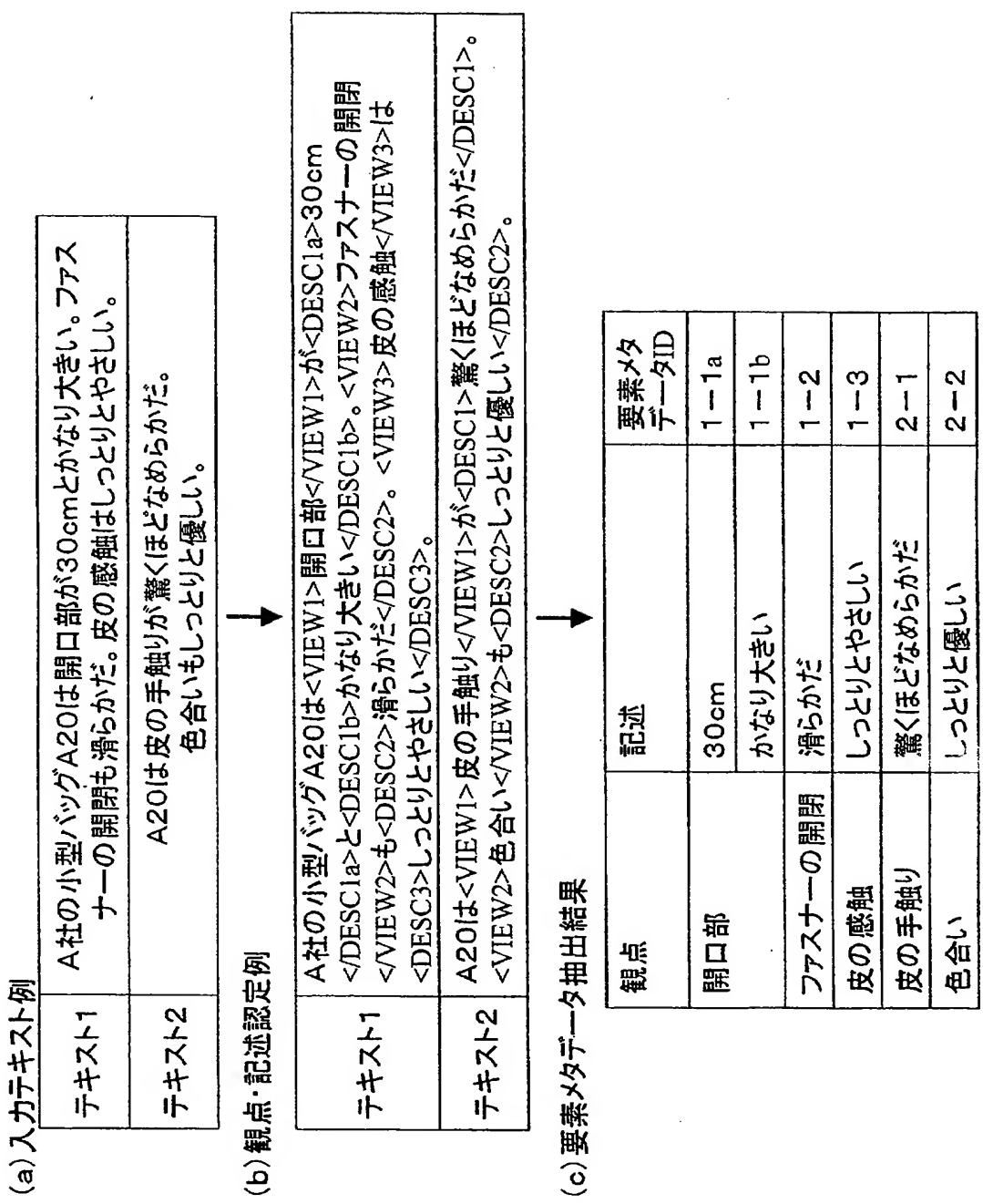
を具備する情報抽出方法。

WO 2005/096182 PCT/JP2005/006317

[図1]



[図2]



(a)入力テキスト例

テキスト1	A社の小型バッグA20は開口部が30cmとかなり大きい。ファスナーの開閉も滑らかだ。皮の感触はしっとりとやさしい。
テキスト2	A20は皮の手触りが驚くほどなめらかだ。色合いもしっとりと優しい。

(b)観点・記述認定例

テキスト1	A社の小型バッグA20は<VIEW1>開口部</VIEW1>が<DESC1a>30cm</DESC1a>と<DESC1b>かなり大きい</DESC1b>。<VIEW2>ファスナーの開閉</VIEW2>も<DESC2>滑らかだ</DESC2>。<VIEW3>皮の感触</VIEW3>は<DESC3>しっとりとやさしい</DESC3>。
テキスト2	A20は<VIEW1>皮の手触り</VIEW1>が<DESC1>驚くほどなめらかだ</DESC1>。<VIEW2>色合い</VIEW2>も<DESC2>しっとりと優しい</DESC2>。

(c)要素メタデータ抽出結果

観点	記述	要素メタデータID
開口部	30cm	1－1a
	かなり大きい	1－1b
ファスナーの開閉	滑らかだ	1－2
皮の感触	しっとりとやさしい	1－3
皮の手触り	驚くほどなめらかだ	2－1
色合い	しっとりと優しい	2－2

[図3]

(a)観点・記述抽出規則例

規則	パタン	観点	記述
1	は({漢字/平仮名連続1})[がも]({英数字連続1})と({漢字/平仮名連続1}{形容詞語尾1})	$1	$2‖$3
2	は({漢字/平仮名連続1})[がも]({漢字/平仮名連続1}{形容動詞語尾1})	$1	$2
3	は({漢字/平仮名連続1})[がも]({漢字/平仮名連続1}{形容詞語尾1})	$1	$2

規則の記法の説明

[ ]:[]内のいずれか

+:直前のパタン要素の1回以上の繰り返し

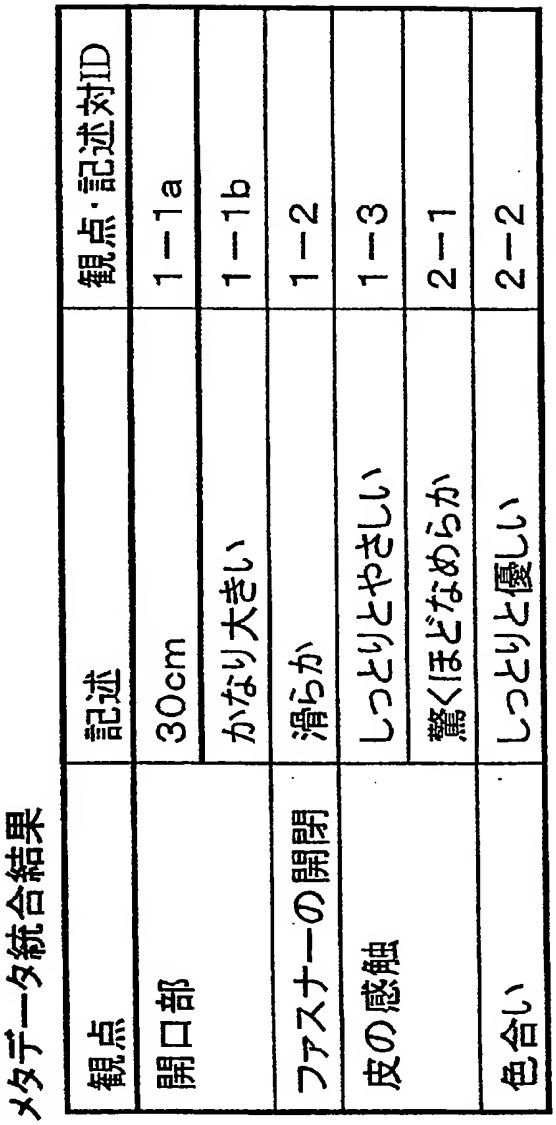
( ):後方参照(順に"${整数}" で参照される)

${整数}:変数(パタンで{整数}番目に"( )"でくくられた部分にマッチする文字列)

(b)観点・記述抽出規則の構成要素定義例

構成要素名	定義
漢字/平仮名連続1	[ぁ-ん亜-瑤]+
英数字連続1	[0-9A-z]+
形容動詞語尾1	だろ,だ,で,だっ,に,な
形容詞語尾1	く,かっ,う,ゆう,い

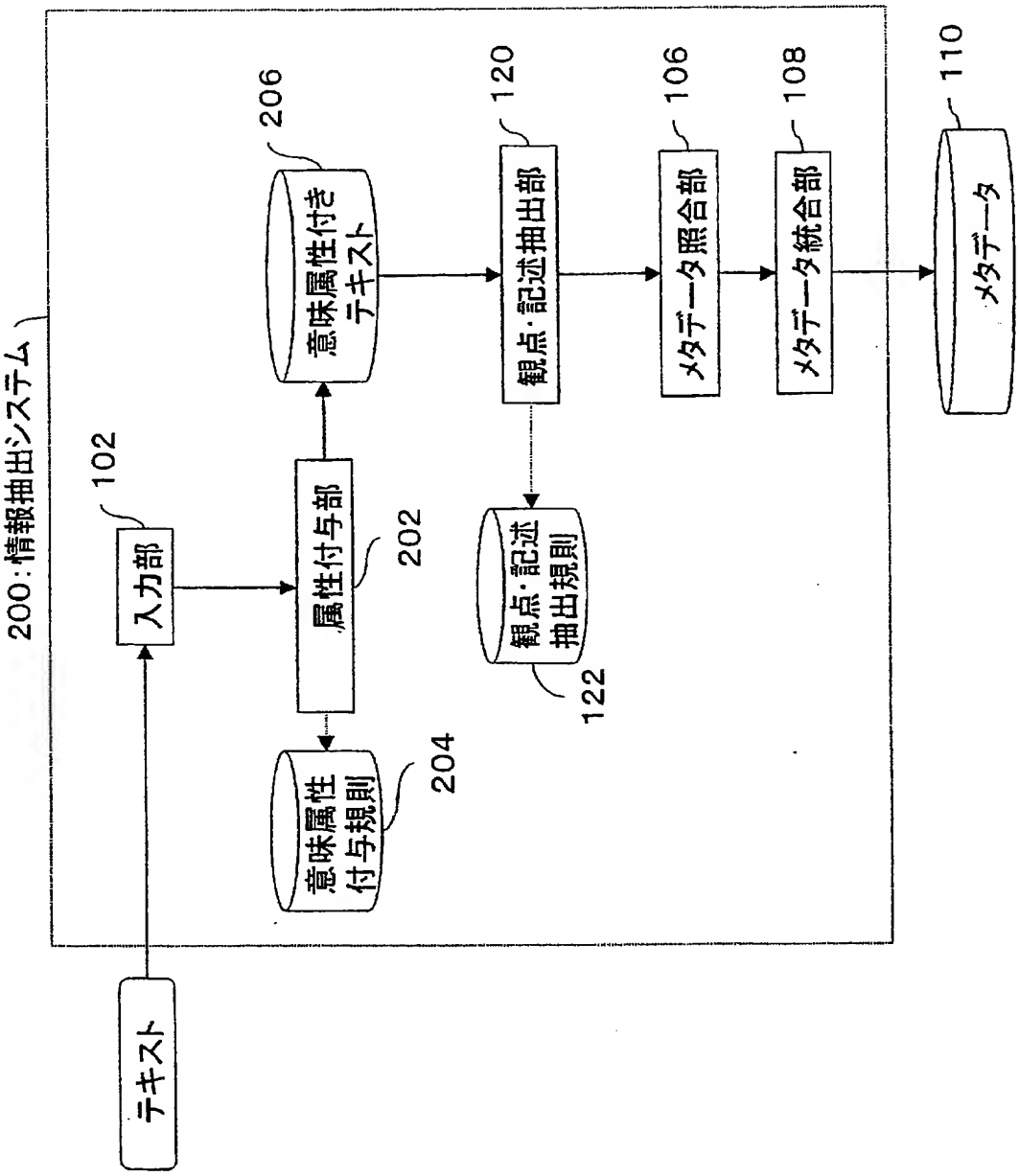
[図4]



メタデータ統合結果

観点	記述	観点・記述対ID
開口部	30cm	1－1a
	かなり大きい	1－1b
ファスナーの開閉	滑らか	1－2
皮の感触	しっとりとやさしい	1－3
	驚くほどなめらか	2－1
色合い	しっとりと優しい	2－2

[図5]


200:情報抽出システム
テキスト
入力部
102
属性付与部
202
意味属性付与規則
204
意味属性付きテキスト
206
観点・記述抽出部
120
観点・記述抽出規則
122
メタデータ照合部
106
メタデータ統合部
108
メタデータ
110

[図6]

(a)テキスト例

テキスト1	A社のバッグA200は容量が20リットルと大きい。
テキスト2	バッグA-200は容量が不十分だと思う。

(b)意味属性付与例

テキスト1	<ORGANIZATION type=company>A社</ORGANIZATION>の<PROD_TYPE>バッグ</PROD_TYPE><PROD_NAME>A200</PROD_NAME>は<QUANT_TYPE>容量</QUANT_TYPE>が<QUANT unit=l, val=20>20リットル</QUANT>と大きい。
テキスト2	<PROD_TYPE>バッグ</PROD_TYPE><PROD_NAME>A-200</PROD_NAME>は<QUANT_TYPE>容量</QUANT_TYPE>が不十分だと思う。

[図7]

(a)意味属性付与規則例

規則	パタン	対象部分	意味属性	
			意味分類	詳細情報
1	({数字連続}{数量単位})	$1 $2	QUANT	・unit:数量単位の表現 ・val:数値連続を正規化した値
2	({数量分類})	$1	QUANT_TYPE	
3	({英字連続}{会社名後接表現})	$1	ORGANIZATION	・type:company
4	({製品分類名})	$1	PROD_TYPE	
5	({製品分類名}{英数字記号連続1})	$1	PROD_NAME	

(b)意味属性付与規則構成要素の定義例

構成要素名	定義
会社名後接表現	'社'
数量分類	'容量'
数量単位	'リットル'、'メートル'、'グラム'、…
製品分類名	'バッグ'、'シューズ'、'帽子'、…
数字連続	[0-9]+
英字連続	[A-Z]+
英数字記号連続1	[−9-9A-z]+

[図8]

(a)意味属性付きテキスト例

テキスト1	<ORGANIZATION type=company>A社</ORGANIZATION>の<PROD_TYPE>バッグ</PROD_TYPE><PROD_NAME>A200</PROD_NAME>は<QUANT_TYPE>容量</QUANT_TYPE>が<QUANT unit=l, val=20>20リットル</QUANT>と大きい。
テキスト2	<PROD_TYPE>バッグ</PROD_TYPE><PROD_NAME>A-200</PROD_NAME>は<QUANT_TYPE>容量</QUANT_TYPE>が不十分だと思う。

↓

(b)観点・記述認定例

テキスト1	<DESC1><ORGANIZATION type=company>A社</ORGANIZATION></DESC1>の<DESC2><PROD_TYPE>バッグ</PROD_TYPE></DESC2> <DESC3><PROD_NAME>A200</PROD_NAME></DESC3>は<VIEW4><QUANT_TYPE>容量</QUANT_TYPE></VIEW4>が<DESC4a><QUANT unit=l, val=20>20リットル</QUANT></DESC4a>と<DESC4b>大きい</DESC4b>。
テキスト2	<DESC1><PROD_TYPE>バッグ</PROD_TYPE></DESC1><DESC2><PROD_NAME>A-200</PROD_NAME></DESC2><VIEW3>は<QUANT_TYPE>容量</QUANT_TYPE></VIEW3>が<DESC3>不十分だ</DESC3>と思う。

[図9]

(a)観点・記述抽出規則例

規則	パタン	観点	記述
1	<QUANT_TYPE>({タグ開始記号以外の任意文字列})</QUANT_TYPE>[がは]<QUANT>({タグ開始記号以外の任意文字列})</QUANT>と({漢字/平仮名連続1}{形容詞語尾1})	$1	$2 ‖ $3
2	<QUANT_TYPE>({タグ開始記号以外の任意文字列})</QUANT_TYPE>[がは]({漢字/平仮名連続}{形容動詞語尾1})	$1	$2
3	{タグ終了記号以外の任意文字}*<({意味属性1})>({タグ開始記号以外の任意文字列})</{意味属性}>	$1の意味属性の別名	$2
4	{終了タグ}<({意味属性1})> <({タグ開始記号以外の任意文字列})> </{意味属性1}>	$1の意味属性の別名	$2

+:直前のパタン要素の1回以上の繰り返し
*:直前のパタン要素の0回以上の繰り返し

(b)観点・記述抽出規則構成要素定義例

構成要素名	定義
形容動詞語尾1	'だろ'、'だ'、'で'、'だっ'、'に'、'な'
形容詞語尾1	'く'、'かっ'、'う'、'ゆう'、'い'
意味属性1	'ORGANIZATION'、'ORGANIZATION type=company' 、'PROD_TYPE'、'PROD_NAME'、 'PERSON'、'DATE'、'TIME'、'PERIOD'、…
タグ開始記号以外の任意文字列1	[^<]+
タグ終了記号以外の任意文字列1	[^>]
終了タグ	</[^>]+>

[図10]

要素メタデータ抽出結果

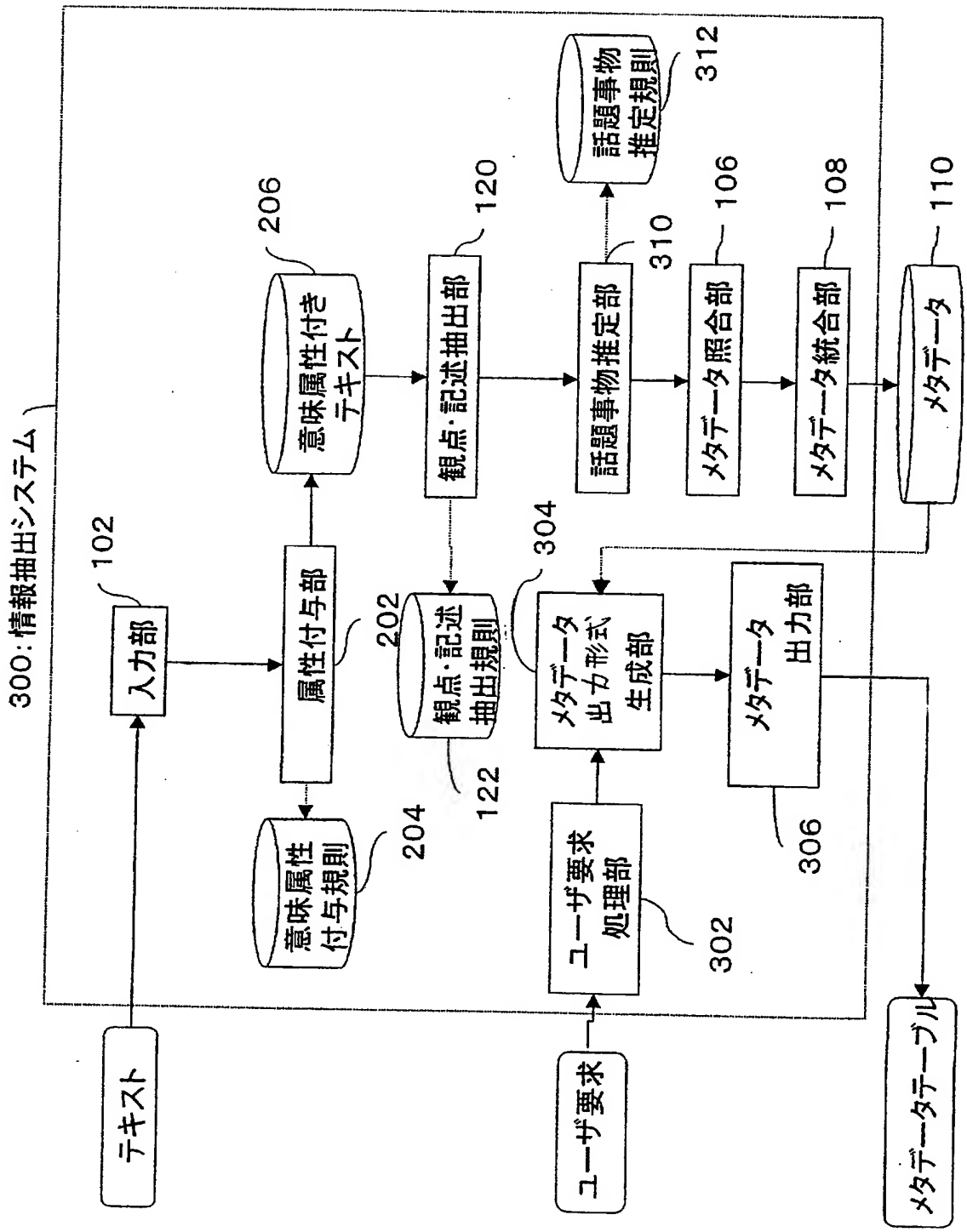
観点	記述	意味属性		要素メタデータID
		意味分類	詳細情報	
会社名	A社	ORGANIZATION	type=company	1－1
製品分類	バッグ	PROD_TYPE		1－2
製品名	A200	PROD_NAME		1－3
容量		QUANT_TYPE		-
	20リットル	QUANT	unit=l, val=20	1－4a
	大きい			1－4b
製品分類	バッグ	PROD_TYPE		2－1
製品名	A-200	PROD_NAME		2－2
容量		QUANT_TYPE		2－3
	不十分だ			

[図11]

メタデータ統合結果

観点	記述	意味属性		要素メタデータID
		意味分類	詳細情報	
会社名	A社	ORGANIZATION	type=company	1−1
製品分類	バッグ	PROD_TYPE		1−2 2−1
製品名	A200	PROD_NAME		1−3 2−2
容量		QUANT_TYPE		-
	20リットル	QUANT	unit=l, val=20	1−4a
	大きい			1−4b
	不十分だ			2−3

[図12]


300:情報抽出システム
テキスト
入力部
102
属性付与部
意味属性付与規則
204
意味属性付き テキスト
206
観点・記述抽出規則
202
観点・記述抽出部
120
122
話題事物推定規則
312
話題事物推定部
310
メタデータ照合部
106
メタデータ統合部
108
メタデータ
110
メタデータ出力形式生成部
304
メタデータ出力部
306
ユーザ要求処理部
302
ユーザ要求
メタデータテーブル

[図13]

(a)観点・記述認定例

テキスト1	<DESC1><PROD_TYPE>バッグ</PROD_TYPE></DESC1><DESC2> <PROD_NAME>A200</PROD_NAME></DESC2>は<VIEW3> <QUANT_TYPE>容量</QUANT_TYPE></VIEW3>が<DESC3>不十分だ</DESC3> し、<DESC4><PROD_TYPE>バッグ</PROD_TYPE></DESC4><DESC5> <PROD_NAME>A300</PROD_NAME></DESC5>は<VIEW6><QUANT_TYPE>容量 </QUANT_TYPE></VIEW6>が<VIEW7>あまりに大きい</VIEW7>。
テキスト2	<DESC1><PROD_TYPE>バッグ</PROD_TYPE></DESC1><DESC2> <PROD_NAME>A200</PROD_NAME></DESC2>は<VIEW3> <QUANT_TYPE>容量</QUANT_TYPE></VIEW3>が<DESC3> <QUANT unit=l val=20>20リットル</QUANT></DESC3>で、 <DESC4><PROD_TYPE>バッグ</PROD_TYPE></DESC4><DESC5> <PROD_NAME>A300</PROD_NAME></DESC5>の<VIEW6><QUANT_TYPE> 容量</QUANT_TYPE></VIEW6>は<VIEW7> <QUANT unit=l val=30>30リットル</QUANT></VIEW7>。

(b)要素メタデータ抽出結果例

観点	記述	意味属性		要素メタデータID
		意味分類	詳細情報	
製品分類	バッグ	PROD_TYPE		1-1
製品名	A200	PROD_NAME		1-2
容量	不十分だ	QUANT_TYPE		1-3
製品分類	バッグ	PROD_TYPE		1-4
製品名	A300	PROD_NAME		1-5
容量	あまりに大きい	QUANT_TYPE		1-6
製品分類	バッグ	PROD_TYPE		2-1
製品名	A200	PROD_NAME		2-2
容量	20リットル	QUANT_TYPE		2-3
		QUANT	unit=l, val=20	
製品分類	バッグ	PROD_TYPE		2-4
製品名	A300	PROD_NAME		2-5
容量	30リットル	QUANT_TYPE		2-6
		QUANT	unit=l, val=30	

[図14]

(a)話題事物推定規則の例

規則	条件	推定される話題事物
1	<DESC[0-9]+><{PROD_TYPE|PERSON}>({タグ開始記号以外の任意文字列}<{PROD_TYPE|PERSON}></DESC[0-9]+>	記述$1の要素メタデータの話題事物は$1
2	({タグ開始記号以外の任意文字列})</{PROD_TYPE|PERSON}></DESC[0-9]+>は(<VIEW[0-9]+><{タグ開始記号以外の任意文字列}></VIEW[0-9]+>が<DESC[0-9]+>{タグ開始記号以外の任意文字列}</DESC[0-9]+>) ただし、$2と$4, $5と$7の文字列は同一であること	観点$3、記述$6の要素メタデータの話題事物は$1
3	({タグ開始記号以外の任意文字列})</{PROD_TYPE|PERSON}></DESC[0-9]+><DESC[0-9]+><{PROD_TYPE|PERSON}><{タグ開始記号以外の任意文字列}></{PROD_TYPE|PERSON}>	記述$2の要素メタデータの話題事物は$4

規則の記法の説明

(A|B):AとBのいずれか

(b)話題事物推定規則構成要素定義の例

構成要素名	定義
タグ構成文字列	[0-9A-Za-z_]+
任意文字列	+

[図15]

話題事物推定例

観点	記述	意味属性		要素メタデータID	話題事物
		意味分類	詳細情報		
製品分類	バッグ	PROD_TYPE		1-1	A200
製品名	A200	PROD_NAME		1-2	A200
		QUANT_TYPE		1-3	A200
容量	不十分だ				
製品分類	バッグ	PROD_TYPE		1-4	A300
製品名	A300	PROD_NAME		1-5	A300
		QUANT_TYPE		1-6	A300
容量	あまりに大きい				
製品分類	バッグ	PROD_TYPE		2-1	A200
製品名	A200	PROD_NAME		2-2	A200
		QUANT_TYPE		2-3	A200
容量	20リットル	QUANT	unit=l, val=20		
製品分類	バッグ	PROD_TYPE		2-4	A300
製品名	A300	PROD_NAME		2-5	A300
		QUANT_TYPE		2-6	A300
容量	30リットル	QUANT	unit=l, val=30		

[図16]

統合結果例

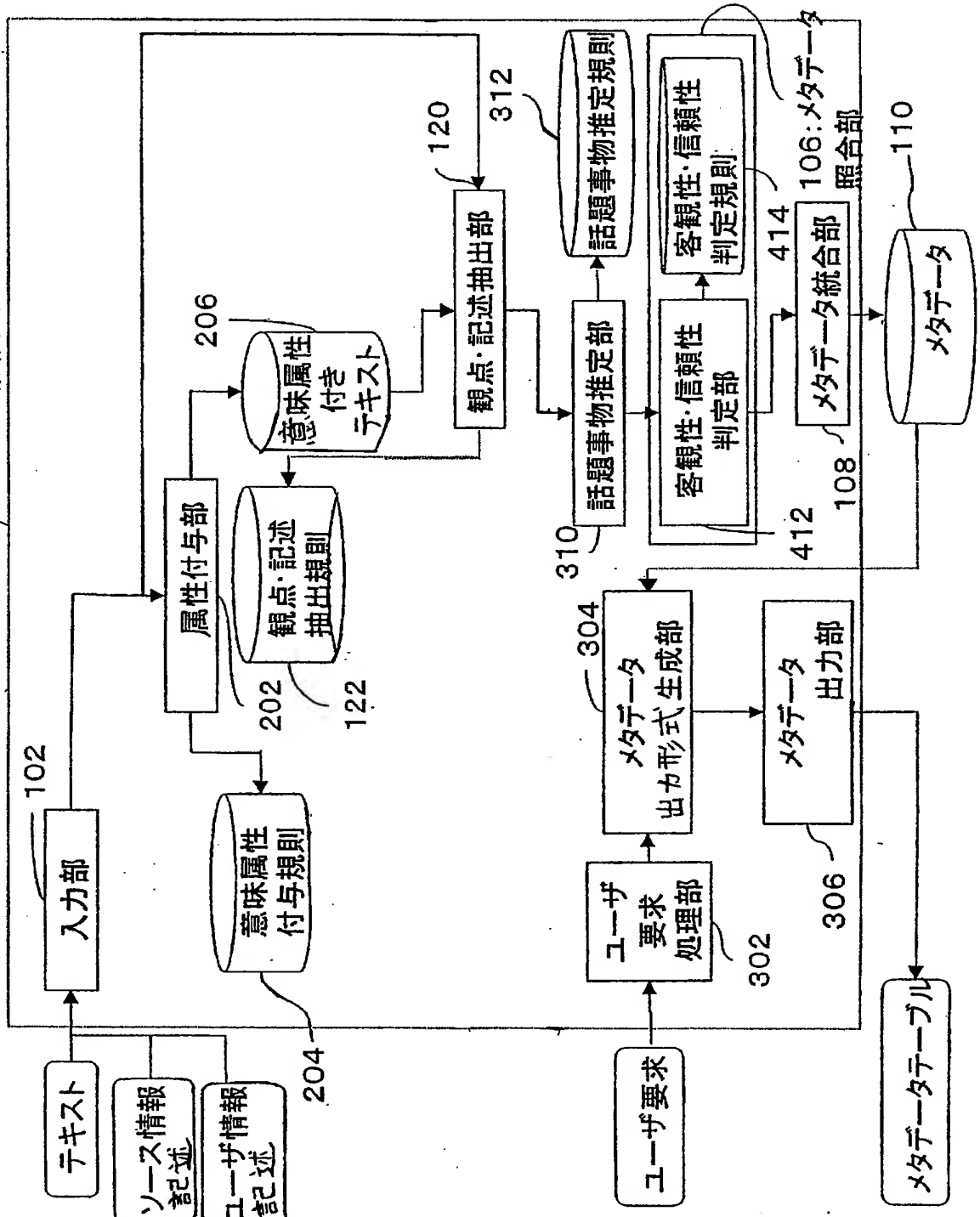
話題事物	観点	記述	意味属性		要素メタデータID
			意味分類	詳細情報	
A200	製品分類	バッグ	PROD_TYPE		1-1 2-1
	製品名	A200	PROD_NAME		1-2 2-2
	容量		QUANT_TYPE		-
		不十分だ			1-3
		20リットル	QUANT	unit=l, val=20	2-3
A300	製品分類	バッグ	PROD_TYPE		1-4 2-4
	製品名	A300	PROD_NAME		1-5 2-5
	容量		QUANT_TYPE		-
		あまりに大きい			1-6
		30リットル	QUANT	unit=l, val=30	2-6

[図17]

メタデータの出力形式の例

話題 事物	観点	記述	意味属性		要素メタデータ ID
			意味分類	詳細情報	
A200	容量		QUANT_TYPE		-
		不十分だ			1-3
		20リットル	QUANT	unit=1 val=20	2-3

[図18]


400:情報抽出システム
テキスト
ソース情報記述
ユーザ情報記述
ユーザ要求
102
入力部
属性付与部
202
意味属性付与規則
204
観点・記述抽出規則
122
206
意味属性付きテキスト
120
観点・記述抽出部
312
話題事物推定規則
310
話題事物推定部
412
客観性・信頼性判定部
414
客観性・信頼性判定規則
106:メタデータ照合部
108
メタデータ統合部
110
メタデータ
304
メタデータ出力形式生成部
306
メタデータ出力部
302
ユーザ要求処理部
メタデータテーブル

[図19]

(a)ソース情報記述の例

ソース情報記述1	http://www.aaa.co.jp/article1 作成日:2003年10月1日
ソース情報記述2	http://www.xxx.yyy.jp/~zzz 作成日:2003年5月1日

(b)ユーザ情報記述の例

ユーザ情報記述1	会社名:aaa
ユーザ情報記述2	作者zzz。20代男性。

(c)意味属性付きソース情報記述の例

ソース情報記述1	<URL type=会社Webページ>http://www.aaa.co.jp/article1</URL> <DATE value=2003:10:01>2003年10月1日</DATE>

(d)意味属性付きユーザ情報の例

ユーザ情報記述1	作者<AUTHOR>zzz</AUTHOR>。<AGE value=20:29 >20代 </AGE><GENDER type=M>男性</GENDER>。

[図20]

(a)ソース意味属性付与規則の例

規則	パタン	意味属性	
		意味分類	詳細情報
1	(http://.*¥.co¥.jp.*)	URL(webページ)	type=company
2	([数字4桁])年([数字1～2桁])月([数字1～2桁])	DATE	value=$1:$2:$3 ただし$2,$3が1桁の場合0を前につける

¥:". "(ピリオド)

(b)ユーザ意味属性付与規則の例

規則	パタン	意味属性	
		意味分類	詳細情報
1	作者[:]*([^.]+)	AUTHOR	
2	([数字1桁])0代	AGE	value=($1*10+ 1): ($1*10+ 9)
3	男性	GENDER	value=M

[^.]:[]内の文字に一致しない文字

[図21]

(a)ソース観点・記述抽出規則の例

規則	パタン	観点	記述
1	<URL type=会社Webページ>({タグ開始記号以外の任意文字列})</URL>	テキストの入手元	$1
2	(作成日):<DATE>({タグ開始記号以外の任意文字列})</DATE>	$1	$2

(b)ユーザ観点・記述抽出規則の例

規則	パタン	観点	記述
1	<AUTHOR>({タグ開始記号以外の任意文字列})</AUTHOR>	作者	$1
2	<AGE value=(20:29)>({タグ開始記号以外の任意文字列})</AGE>	年齢	$1
3	<GENDER type=M>({タグ開始記号以外の任意文字列})</GENDER>	性別	男性

[図22]

(a)ソースメタデータ抽出結果の例

観点	記述	意味属性		ソースメタデータID
		意味分類	詳細情報	
テキストの入手元	http://www.aaa.co.jp/article1	URL(webページ)	type=company	1−S1
テキストの分類	会社webページ (http://www.aaa.co.jp/article1)	URL(webページ)	type=company	1−S2
作成日	2003年10月1日	DATE	value=2003:10:01	1−S3

(b)ユーザメタデータ抽出結果の例

観点	記述	意味属性		ユーザメタデータID
		意味分類	詳細情報	
作者	zzz	AUTHOR		2−U1
年齢	20-29	AGE	value=20:29	2−U2
性別	男性	GENDER	value=M	2−U3

[図23]

客観性・信頼性判定規則例

規則	条件					客観性	信頼性
	テキストの要素メタデータ		ソースメタデータ		その他		
	観点	記述の意味分類	観点	記述			
1	組織名					1	1
2	製品分類					1	1
3	製品名					1	1
4	用途	USAGE	テキストの分類	会社webページ/新聞記事		1	1
5	用途	USAGE	テキストの分類	個人webページ		0	0.5
6	容量	QUANT	テキストの分類	会社webページ/新聞記事		1	1
7	容量	QUANT	テキストの分類	個人webページ		1	0.9
8	容量	なし、またはQUANT以外			文末が{不確定表現1}	0	0.2
9	容量	なし、またはQUANT以外	テキストの分類	会社webページ/新聞記事	文末が{不確定表現1}以外	0	0.5
10	容量	なし、またはQUANT以外	テキストの分類	個人webページ	文末が{不確定表現1}以外	0	0.3

客観性・信頼性判定規則構成要素定義例

不確定表現1:と思う,と思える,と思われる,かもしれない,らしい

[図24]

(a)テキスト例

テキスト1	バッグA200は容量が20リットルと大きい。
テキスト2	バッグA200の容量は海外出張用には不十分だと思う。
テキスト3	バッグA200の容量は国内出張用にはあまりに大きい。
テキスト4	バッグA200の容量は国内出張用には十分だ。

(b)意味属性付きテキスト例

テキスト1	<ORGANIZATION type=company>A社</ORGANIZATION>の<PROD_TYPE>バッグ</PROD_TYPE><PROD_NAME>A200</PROD_NAME>は<QUANT_TYPE>容量</QUANT_TYPE>が<QUANT unit=l, val=20>20リットル</QUANT>と大きい。
テキスト2	<PROD_TYPE>バッグ</PROD_TYPE><PROD_NAME>A200<PROD_NAME>の<QUANT_TYPE>容量</QUANT_TYPE>は<USAGE>海外出張用</USAGE>には不十分だと思う。
テキスト3	<PROD_TYPE>バッグ</PROD_TYPE><PROD_NAME>A200<PROD_NAME>の<QUANT_TYPE>容量</QUANT_TYPE>は<USAGE>国内出張用</USAGE>にはあまりに大きい。
テキスト4	<PROD_TYPE>バッグ</PROD_TYPE><PROD_NAME>A200<PROD_NAME>の<QUANT_TYPE>容量</QUANT_TYPE>は<USAGE>国内出張用</USAGE>には十分だ。

[図25]

(a)観点・記述抽出規則例

規則	パタン	観点	記述
1	<QUANT_TYPE>(<タグ開始記号以外の任意文字列>)(</QUANT_TYPE>)[が|は]<QUANT>(<タグ開始記号以外の任意文字列>)</QUANT>と({漢字/平仮名連続1}{形容詞語尾1})	$1	$2||$3
2	<QUANT_TYPE>(<タグ開始記号以外の任意文字列>)(</QUANT_TYPE>)[が|は]<USAGE>(<タグ開始記号以外の任意文字列>用)</USAGE>(に|として)[はも]({漢字/平仮名連続1}{形容詞語尾1})	$1	$2&&$3
3	<QUANT_TYPE>(<タグ開始記号以外の任意文字列>)</QUANT_TYPE>[が|は]({漢字/平仮名連続1}{形容動詞語尾1})	$1	$2
4	{タグ終了記号以外の任意文字列}*<({意味属性})>({タグ開始記号以外の任意文字列})</{意味属性1}>	$1の意味属性の別名	$2
5	{終了タグ}<({意味属性1})> <({タグ開始記号以外の任意文字列})></{意味属性1}>	$1の意味属性の別名	$2

(b)観点・記述抽出規則構成要素定義例

構成要素名	定義
タグ終了記号以外の任意文字	[^>]
終了タグ	</[^>]+>

[図26]

要素メタデータ抽出結果例

観点	記述	意味属性		要素メタデータID	話題事物
		意味分類	詳細情報		
製品分類	バッグ	PROD_TYPE		1−1	A-200
製品名	A200	PROD_NAME		1−2	A-200
容量		QUANT_TYPE		-	A-200
	20リットル	QUANT	unit=l, val=20	1−3a	
	大きい	-		1−3b	
製品分類	バッグ	PROD_TYPE		2−1	A-200
製品名	A200	PROD_NAME		2−2	A-200
容量		QUANT_TYPE		2−3	A-200
	海外出張用	USAGE			
	不十分だ	-			
用途	海外出張用	USAGE		2−4	A-200
製品分類	バッグ	PROD_TYPE		3−1	A-200
製品名	A200	PROD_NAME		3−2	A-200
容量		QUANT_TYPE		3−3	A-200
	国内出張用	USAGE			
	あまりに大きい	-			
用途	国内出張用	USAGE		3−4	A-200
製品分類	バッグ	PROD_TYPE		4−1	A-200
製品名	A200	PROD_NAME		4−2	A-200
		QUANT_TYPE		4−3	A-200
	国内出張用	USAGE			
容量	十分だ	-			
用途	国内出張用	USAGE		4−4	A-200

[図27]

容観性·信頼性判定結果例

要素メタデータ						ソースメタデータ			ユーザメタデータ		
観点	記述	意味属性	要素メタデータID	容観性	信頼性	観点	記述	ソースメタデータID	観点	記述	ユーザメタデータID
製品分類	バッグ	PROD_TYPE	1−1	1.0	1.0	テキストの分類	会社webページ	1−S2	性別	-	-
製品名	A200	PROD_NAME	1−2	1.0	1.0						
容量		QUANT_TYPE	-	-	-						
	20リットル	QUANT	1−3a	1.0	1.0						
	大きい	-	1−3b	0	0.5						
製品分類	バッグ	PROD_TYPE	2−1	1.0	1.0		個人webページ	2−S2		男性	2−U3
製品名	A200	PROD_NAME	2−2	1.0	1.0						
容量		QUANT_TYPE	-	-							
	海外出張用	USAGE	2−3	0	0.2						
	不十分だ	-									
用途	海外出張用	USAGE	2−4	0	0.5						
製品分類	バッグ	PROD_TYPE	3−1	1.0	1.0		個人webページ	3−S2		女性	3−U3
製品名	A200	PROD_NAME	3−2	1.0	1.0						
容量		QUANT_TYPE	-								
	国内出張用	USAGE	3−3	0	0.3						
	あまりに大きい	-									
用途	国内出張用	USAGE	3−4	0	0.5						
製品分類	バッグ	PROD_TYPE	4−1	1.0	1.0		個人webページ	4−S2		男性	4−U3
製品名	A200	PROD_NAME	4−2	1.0	1.0						
容量		QUANT_TYPE	-								
	国内出張用	USAGE	4−3	0	0.3						
	十分だ	-									
用途	国内出張用	USAGE	4−4	0	0.5						

[図28]

メタデータ統合結果例

要素メタデータ							ソースメタデータ			ユーザメタデータ		
話題事物	観点	記述	意味属性	要素メタデータID	客観性	信頼性	観点	記述	ソースメタデータID	観点	記述	ユーザメタデータID
A200	製品分類	バッグ	PROD_TYPE	1−1	1.0	1.0	テキストの分類	会社webページ	1−S2	性別	-	-
				2−1				個人webページ	2−S2		男性	2−U3
				3−1					3−S2		女性	3−U3
				4−1					4−S2		男性	4−U3
	製品名	A200	PROD_NAME	1−2	1.0	1.0		会社webページ	1−S2		-	
				2−2				個人webページ	2−S2		男性	2−U3
				3−2					3−S2		女性	3−U3
				4−2					4−S2		男性	4−U3
	容量		QUANT_TYPE	-							-	-
		20リットル	QUANT	1−3a	1.0	1.0		会社webページ	1−S2		-	-
		大きい	-	1−3b	0	0.5						
		海外出張用	USAGE	2−3	0	0.2		個人webページ	2−S2		男性	2−U3
		不十分だ	-									
		国内出張用	USAGE	3−3	0	0.3			3−S2		女性	3−U3
		あまりに大きい	-									
		国内出張用	USAGE	4−3	0	0.3			4−S2		男性	4−U3
		十分だ	-									
	用途	海外出張用	USAGE	2−4	0	0.5		個人webページ	2−S2		男性	2−U3
		国内出張用	USAGE	3−4					3−S2		女性	3−U3
				4−4					4−S2		男性	4−U3

[図29]

メタデータテーブル

要素メタデータ							ソースメタデータ			ユーザメタデータ		
話題事物	観点	記述	意味属性	要素メタデータID	客観性	信頼性	観点	記述	ソースメタデータID	観点	記述	ユーザメタデータID
A200	容量		QUANT_TYPE	-	-	-	テキストの種類	個人webページ		性別	-	
		海外出張用	USAGE	2−3	0	0.2			2−S2		男性	2−U3
		不十分だ	-									
		国内出張用	USAGE	3−3	0	0.3			3−S2		女性	3−U3
		あまりに大きい	-									
		国内出張用	USAGE	4−3	0	0.3			4−S2		男性	4−U3
		十分だ	-									
	用途	海外出張用	USAGE	2−4	0	0.5			2−S2		男性	2−U3
		国内出張用	USAGE	3−4					3−S2		女性	3−U3
				4−4					4−S2		男性	4−U3